Rappresentazione, previsione e mondo interno: un'indagine su JEPA e World Models sulla strada verso l'Intelligenza Artificiale Generale (AGI)

Abstract

La presente monografia indaga le architetture JEPA (*Joint-Embedding Predictive Architecture*) e i *World Models* come risposta teoricamente fondata ai limiti strutturali del paradigma generativo-predittivo dominante nell'apprendimento automatico contemporaneo. La tesi centrale è che i grandi modelli attuali, pur mostrando straordinarie capacità empiriche, soffrono di un'inadeguatezza di principio: apprendendo distribuzioni sui segnali sensoriali grezzi anziché strutture causali e relazionali, acquisiscono competenza funzionale senza comprensione genuina. JEPA, spostando il bersaglio dell'apprendimento dalla predizione di osservazioni grezze alla predizione di rappresentazioni latenti, offre un'alternativa architetturale di rilievo. L'analisi si articola su tre piani intrecciati — tecnico-matematico, storico-genealogico e filosofico-critico — e copre l'intera traiettoria dal formalismo di I-JEPA e V-JEPA al framework *Autonomous Machine Intelligence* di LeCun, dai *World Models* di Schmidhuber e Hafner alle estensioni teoriche di LeJEPA e *Semantic Tube Prediction*. I capitoli conclusivi problematizzano le pretese del paradigma computazionale attraverso i teoremi di impossibilità di Putnam, Kripke e Tarski-Gödel, formulando un giudizio differenziato sulla plausibilità del percorso JEPA→AMI→AGI e sui suoi limiti strutturali irriducibili.

Introduzione

Questa monografia nasce da una constatazione semplice nella sua formulazione e perturbante nelle sue implicazioni: i sistemi di intelligenza artificiale più potenti che l'umanità abbia mai prodotto sembrano sapere moltissimo e capire pochissimo. Non si tratta di un'iperbole retorica. I grandi modelli linguistici contemporanei generano testi di straordinaria coerenza sintattica e semantica su qualsiasi argomento; i modelli visivi classificano oggetti con precisione superiore a quella umana in condizioni controllate; i sistemi multimodali descrivono scene, rispondono a domande, producono codice funzionante. Eppure questi stessi sistemi falliscono in modo bizzarro e sistematico di fronte a variazioni minime dei compiti, a riformulazioni innocue delle domande, a combinazioni di concetti mai incontrate nell'addestramento. Questa coesistenza di straordinaria capacità e strutturale fragilità non è un difetto contingente correggibile con più dati o più parametri: è il sintomo di un'inadeguatezza profonda nell'obiettivo stesso dell'apprendimento. Ed è precisamente questa diagnosi che ha motivato lo sviluppo delle architetture analizzate nel presente lavoro.

Il paradigma dominante dell'apprendimento automatico degli ultimi quindici anni — che questa monografia chiama paradigma generativo-predittivo a livello sensoriale — ha come obiettivo la ricostruzione o la predizione dell'input grezzo: predire il prossimo token di testo, ricostruire una regione mascherata di un'immagine, generare un frame video coerente con i precedenti. Tale obiettivo ha prodotto risultati empirici di enorme valore pratico. Ma contiene una distorsione strutturale che i capitoli di questo paper analizzano con rigore crescente: costringendo il modello a operare nello spazio delle osservazioni sensoriali — pixel, campioni audio, sequenze di token discreti — lo si obbliga a dedicare capacità rappresentativa alla modellazione di varianti percettive semanticamente irrilevanti (il preciso pattern di rumore in un'immagine, le fluttuazioni di fase di un segnale audio, la microtexture di una superficie) a scapito delle strutture causali, relazionali e composizionali che costituiscono il nucleo della comprensione del mondo. JEPA — la famiglia di architetture a Joint-Embedding Predictive Architecture proposta da Yann LeCun e dai suoi collaboratori a partire dal 2022 — risponde a questa diagnosi spostando il bersaglio dell'apprendimento: invece di predire osservazioni grezze, il sistema apprende a predire rappresentazioni latenti. La differenza, apparentemente tecnica, ha conseguenze teoriche e pratiche che in questa sede ci si propone di esplorare nella loro interezza.

La monografia che il lettore ha tra le mani affronta JEPA e i sistemi correlati — i World Models nella tradizione da Schmidhuber a Hafner, il framework dell'Autonomous Machine Intelligence di LeCun, le recenti estensioni teoriche di LeJEPA e della Semantic Tube Prediction — attraverso tre piani simultanei e deliberatamente intrecciati. Il primo piano è quello tecnico-matematico: le architetture vengono presentate con il grado di formalità necessario a rendere precise le intuizioni, a distinguere ciò che è stato dimostrato da ciò che è stato congetturato, a valutare la solidità delle giustificazioni teoriche rispetto ai risultati empirici. Il secondo piano è quello storico e genealogico: ogni architettura viene collocata nella tradizione intellettuale da cui proviene, mostrando le linee di continuità che la connettono ai precursori — dalla predictive coding di Helmholtz e Rao-Ballard all'epistemologia costruttivista di Piaget, dai World Models di Schmidhuber agli ELBO variazionali di Hafner — e le rotture che la distinguono. Il terzo piano è quello filosofico-critico: le domande che la tecnica non può rispondere da sola — cosa significa comprensione nel senso rilevante per l'AGI, se la sintassi è sufficiente per la semantica, dove si trovano i limiti strutturali del paradigma computazionale — vengono affrontate con la stessa serietà intellettuale riservata alle derivazioni matematiche.

La struttura dell'opera riflette questa triplice articolazione. I primi due capitoli costruiscono le fondamenta: il Capitolo I diagnostica il problema del paradigma generativo-predittivo attraverso l'analisi epistemologica dei suoi limiti strutturali, collocando la questione della rappresentazione nel più ampio contesto della filosofia della conoscenza da Piaget a Helmholtz a LeCun; il Capitolo II formalizza gli strumenti matematici necessari — spazi delle rappresentazioni, teoria dell'informazione, formalismo variazionale, geometria della loss landscape —

senza i quali le intuizioni successive resterebbero imprecise. I Capitoli III e IV sono il cuore tecnico della monografia: il Capitolo III descrive l'architettura I-JEPA e le sue estensioni (V-JEPA, LeJEPA, la Semantic Tube Prediction, il sistema DAAM per il parlato) con piena formalità matematica; il Capitolo IV sviluppa la storia e il formalismo dei World Models, dall'architettura di Ha e Schmidhuber (2018) all'RSSM di Hafner fino a DreamerV3 e GAIA-1. Il Capitolo V integra queste architetture nel framework AMI di LeCun, analizzando i sei moduli funzionali del sistema, il principio del Model Predictive Control come meccanismo di pianificazione e la struttura degli Energy-Based Models come linguaggio dell'interazione modulare. I Capitoli VI, VII e VIII affrontano rispettivamente la fattibilità computazionale e la sostenibilità energetica del percorso verso l'AGI, l'estensione del principio JEPA al dominio del parlato, e le implicazioni multimodali del framework per un Perception Module universale. I Capitoli IX e X si spostano deliberatamente sul piano filosofico-critico: il Capitolo IX analizza le tensioni tra comprensione tecnica e comprensione genuina attraverso Searle, Chalmers, Dennett, Harnad e Pearl; il Capitolo X formula un giudizio differenziato sulla plausibilità della via JEPA→AMI→AGI distinguendo ciò che è stato dimostrato da ciò che rimane speculativo. Il Capitolo XI, infine, radicalizza la critica attraverso tre teoremi di impossibilità — Putnam, Kripke, Tarski-Gödel — e il meccanismo del lavaggio epistemico, diagnosticando i limiti strutturali del paradigma meccanicistico che JEPA, come qualsiasi sistema computazionale, eredita per costituzione.

Una nota sul destinatario presupposto e sullo stile. Questo testo è scritto per chi non si accontenta di capire "come funziona" nel senso del ricercatore che vuole replicare un esperimento, né di capire "cosa significa" nel senso del filosofo che non intende 'sporcarsi le mani' con i dettagli tecnici. L'ideale che ha guidato la stesura è quello del lettore che vuole capire *perché*: perché certe architetture sono più promettenti di altre, perché certi problemi resistono a qualsiasi aumento di scala, perché la distinzione tra competenza funzionale e comprensione genuina non è filosoficamente oziosa ma ha conseguenze ingegneristiche precise. Le derivazioni matematiche — dalla perdita di I-JEPA all'analisi spettrale dell'attenzione, dall'ELBO dell'RSSM alle dimostrazioni di impossibilità nel Capitolo XI — sono parte integrante della narrazione, non ornamento tecnico separabile. Esse vengono presentate non per imporre rigore ma perché in questo campo, come si argomenta esplicitamente nella sezione I.7.A, la matematica non è il linguaggio con cui si esprimono intuizioni già possedute: è lo strumento attraverso il quale le intuizioni diventano abbastanza precise da essere verificate, falsificate o generalizzate. Il lettore non matematicamente inclinato può seguire la narrativa concettuale e tornare alle derivazioni più tecniche in seconda battuta; ma chi leggerà la matematica troverà connessioni che la sola narrazione non avrebbe potuto rivelare.

Sul piano delle fonti: la letteratura di riferimento principale è quella pubblicata tra il 2018 e il 2026, un arco di tempo in cui il panorama dell'apprendimento autosupervisionato è mutato con una velocità che rende qualsiasi sintesi parziale

per costruzione. Tra i lavori più recenti, tre meritano una menzione esplicita già in questa sede introduttiva: il documento programmatico "A Path Towards Autonomous Machine Intelligence" di LeCun (2022), che costituisce il manifesto intellettuale dell'intera architettura AMI e è analizzato in dettaglio nel Capitolo V; il paper "LeJEPA" di Balestriero e LeCun (2025), che per la prima volta deriva rigorosamente la distribuzione ottimale degli embedding per qualsiasi compito di probing, trasformando JEPA da architettura empiricamente vincente a sistema teoricamente fondato; e il paper "Semantic Tube Prediction" di Huang, LeCun e Balestriero (2026), che dimostra empiricamente la violazione del termine dati delle Scaling Laws in stile Chinchilla attraverso la regolarizzazione geodetica delle traiettorie latenti. Un quarto lavoro, analizzato nelle sezioni integrative del Capitolo III, completa questo nucleo teorico: il paper "Gaussian Joint Embeddings" di Huang (2026, arXiv:2603.26799), che formalizza JEPA come caso speciale di una modellazione generativa della densità congiunta p(z_c, z_t), dimostra algebricamente il Mahalanobis Trace Trap come causa strutturale del collasso, e prova che l'apprendimento contrastivo standard (SimCLR, MoCo, InfoNCE) è un caso limite degenere del framework. Un quinto lavoro, analizzato nella sezione III.Y, estende il quadro verso il regime sparso: il paper "Rectified LpJEPA" di Kuang, Dagade, Rudner, Balestriero e LeCun (2026, arXiv:2602.01456), che introduce la Rectified Distribution Matching Regularization (RDMReg) e la famiglia delle distribuzioni Gaussiane Generalizzate Rettificate (RGG), dimostrando che sparsità ℓ0 controllabile e massima entropia — riscalata dalla dimensione di informazione di Rényi — non sono obiettivi in conflitto. Questi cinque lavori, insieme al corpus di risultati su I-JEPA, V-JEPA e V-JEPA 2, costituiscono il nucleo empirico e teorico su cui l'argomentazione si appoggia. La critica filosofica del Capitolo XI integra questa trattazione con le opere di Putnam, Kripke, Tarski, Searle e Wittgenstein, mostrando che i limiti strutturali del paradigma computazionale non sono contingenti ma dimostrabili per via logico-matematica.

Ciò che questa monografia non è: non è una rassegna neutrale della letteratura; non è un manuale operativo per implementare architetture JEPA; non è un testo di divulgazione che semplifica per rendere accessibile. È invece un tentativo di fare epistemologia dell'intelligenza artificiale: di applicare alla ricerca sull'IA gli stessi standard di rigore, onestà e autocritica che quella ricerca dovrebbe applicare ai problemi che studia. Il giudizio finale sul framework JEPA→AMI→AGI — espresso nella sezione X.4 — non è né l'ottimismo acritico dei comunicati stampa né lo scetticismo a priori di chi nega l'importanza di questi sviluppi. È un giudizio differenziato, che distingue livelli di plausibilità diversi per componenti diversi del framework, e che riconosce con serietà dove si trovano le barriere fondamentali ancora irrisolte. Se questa monografia riesce a trasmettere quella combinazione di rigore tecnico, onestà epistemica e sensibilità filosofica che considera necessaria per affrontare l'intelligenza artificiale con la serietà che merita, avrà raggiunto il suo scopo principale.

Capitolo I — Prologo Epistemologico: La Crisi del Paradigma Generativo e il Problema Fondamentale della Rappresentazione

I.1 — Il paradosso della conoscenza senza comprensione

Vi è una tensione costitutiva al cuore dell'intelligenza artificiale contemporanea che non può essere sciolta attraverso mere raffinazioni ingegneristiche: i sistemi più capaci che l'umanità abbia mai costruito producono output di straordinaria qualità pur restando, in un senso preciso e non metaforico, privi di comprensione. Questa affermazione non è un lamento filosofico di maniera né un'eco del dibattito sulla coscienza delle macchine — argomento che qui si sfiora ma non si affronta direttamente. È invece una diagnosi tecnica che emerge dall'analisi delle architetture dominanti, dei loro meccanismi di addestramento e delle loro sistematiche modalità di fallimento.

Il paradigma che ha dominato l'ultimo decennio dell'apprendimento automatico — quello che potremmo chiamare paradigma generativo-predittivo a livello sensoriale — si è dimostrato straordinariamente fecondo sul piano empirico e tuttavia strutturalmente limitato sul piano cognitivo. La domanda che anima questa monografia è: perché? E, conseguentemente, quali architetture alternative potrebbero superare tale limitazione in modo non accidentale ma per ragioni di principio?

Per rispondere con rigore bisogna risalire alla questione della rappresentazione, che è al tempo stesso il problema più antico dell'epistemologia e il problema più urgente dell'intelligenza artificiale moderna. Una rappresentazione è, nel senso più astratto, una funzione $\varphi: \mathcal{X} \to \mathcal{Z}$ che mappa elementi di uno spazio di input grezzo $\mathcal{X}$ — immagini, suoni, testo, sequenze di azioni — in elementi di uno spazio latente $\mathcal{Z}$ presumibilmente dotato di struttura più favorevole per il ragionamento, la pianificazione e la generalizzazione. Il nodo cruciale non è l'esistenza di tale funzione, che è sempre garantita in senso banale, ma la sua qualità: una buona rappresentazione dovrebbe collassare insieme varianti irrilevanti dello stesso concetto, separare concetti distinti, catturare relazioni causali, supportare il trasferimento ad ambienti nuovi e permettere la composizione sistematica di concetti elementari in strutture complesse.

L'apprendimento profondo ha imparato a costruire rappresentazioni di qualità sorprendente per compiti specifici, ma il percorso verso rappresentazioni genuinamente generali rimane aperto. La radice del problema non sta nella quantità di parametri o dati — ipotesi che si è ripetutamente dimostrata insufficiente — ma nella natura stessa degli obiettivi di addestramento. E questo è precisamente il punto di partenza teorico di JEPA.

I.1.A — Dalla crisi del paradigma all'agenda di ricerca: perché JEPA

Per comprendere la profondità della tensione tra capacità e comprensione, è utile esaminarla attraverso la lente dei benchmark che la comunità dell'apprendimento

automatico usa per misurare il progresso. BIG-Bench, pubblicato nel 2022 con oltre duecento compiti linguistici, ha documentato sistematicamente il fenomeno che i suoi autori hanno chiamato "generalizzazione apparente": i Large Language Models mostrano prestazioni che crescono con la scala del modello su compiti che sembrano richiedere ragionamento, ma che crollano non appena i compiti vengono formulati in modo leggermente diverso da come appaiono nel pre-training. Un modello che risolve correttamente l'80% dei problemi aritmetici formulati in linguaggio naturale standard può collassare al 20% quando gli stessi problemi vengono presentati con formattazione inusuale — modifiche irrilevanti per il contenuto matematico ma rilevanti per la distribuzione dei token nel corpus di addestramento.

Questo fenomeno non è un difetto di ingegneria correggibile con più parametri: è la conseguenza strutturale del fatto che i modelli apprendono distribuzioni condizionali sui token, non strutture causali. Un modello che apprende p(x_{t+1} | x_{≤t}) ottimizzato su testi matematici impara che dopo "2 + 2 =" segue spesso "4", non che l'addizione di due unità produce quattro. La prima è una relazione statistica tra sequenze di simboli; la seconda è una relazione strutturale nel mondo. Confondere le due è l'errore concettuale che JEPA cerca di aggirare spostando l'obiettivo di apprendimento dallo spazio dei simboli allo spazio delle strutture.

I.2 — Il regime del pixel e i suoi limiti strutturali

Comprendere JEPA richiede di comprendere innanzitutto ciò che JEPA cerca di superare. Il paradigma dominante nell'apprendimento autosupervisionato delle immagini — e in larga misura nel trattamento di qualsiasi segnale ad alta dimensionalità — è stato per anni quello della predizione o ricostruzione a livello di pixel, o più in generale a livello di rappresentazione sensoriale grezza.

L'idea fondante è di natura compressiva: se un modello riesce a ricostruire una regione mascherata di un'immagine partendo dal contesto non mascherato, allora deve aver acquisito una rappresentazione interna del mondo sufficientemente ricca da supportare quella ricostruzione. Questo ragionamento, pur non privo di merito, contiene un difetto concettuale che diventa sempre più evidente man mano che lo si analizza con attenzione.

Consideriamo formalmente il problema. Sia $\mathbf{x} \in \mathbb{R}^d$ un'immagine ad alta risoluzione, sia $\mathbf{x}_C$ la porzione di essa visibile come contesto e $\mathbf{x}_T$ la porzione mascherata che il modello deve predire. Un modello MAE (Masked Autoencoder) o un modello simile apprende una funzione:

$$f_\theta: \mathbb{R}^{|C|} \to \mathbb{R}^{|T|}$$

che minimizza una perdita di ricostruzione del tipo:

$$\mathcal{L}_{rec}(\theta) = \mathbb{E}_{\mathbf{x}} \left[\| \mathbf{x}_T - f_\theta(\mathbf{x}_C) \|^2 \right]$$

oppure, nel caso di modelli di diffusione condizionata, una perdita più sofisticata sulla distribuzione condizionale p(**x**_T | **x**_C). Il punto critico è che in entrambi i casi l'obiettivo è definito nello spazio dei pixel grezzi, ovvero nello spazio $\mathbb{R}^{|T|}$ dove |T| ha tipicamente dell'ordine di grandezza di 10^4–10^5 dimensioni per immagini di risoluzione moderna.

Il problema di una tale formulazione è duplice. In primo luogo, la distribuzione p(**x**_T | **x**_C) è multimodale in modo estremo: dato il contesto di un'immagine raffigurante un volto umano con la metà superiore mascherata, vi sono migliaia di completamenti plausibili che differiscono nei dettagli di texture, colore preciso, illuminazione, microstruttura della pelle — tutti ugualmente compatibili con la semantica della scena. Un modello deterministico che minimizza la perdita quadratica media sarà pertanto costretto a predire la media di questa distribuzione, che corrisponde a un'immagine sfumata priva di dettaglio ad alta frequenza. Un modello che apprende la distribuzione intera — come i modelli di diffusione — può campionare da essa ma deve necessariamente dedicare grandissima capacità rappresentativa alla modellazione di varianze semanticamente irrilevanti: la precisa configurazione del capillare in una guancia, la microtrama di un muro sullo sfondo, il pattern casuale del rumore di sensore.

In secondo luogo, e più sottilmente, il fatto che la predizione avvenga nello spazio sensoriale introduce un bias architetturale che tende a premiare la cattura di correlazioni locali a bassa frequenza spaziale a scapito delle relazioni semantiche a lungo raggio. Una rete convoluzionale — ma in misura minore anche un Vision Transformer — che deve predire pixel tende a imparare filtri di texture e completamento locale piuttosto che relazioni di tipo "questa regione è un occhio umano e quella regione dovrebbe contenere un altro occhio simmetrico a circa tale distanza". Le proprietà globali e semantiche emergono solo come effetti collaterali dell'ottimizzazione locale, non come conseguenza diretta dell'obiettivo.

Yann LeCun ha articolato questa critica in modo particolarmente incisivo nel suo documento programmatico del 2022, "A Path Towards Autonomous Machine Intelligence", dove osserva che i modelli generativi a livello sensoriale sono fondamentalmente mal posti per l'apprendimento di rappresentazioni del mondo utili per la pianificazione e il ragionamento. La ragione di fondo è che il mondo, nel senso in cui è rilevante per un agente che deve agire in esso, non è un segnale ad alta dimensionalità da ricostruire nei suoi dettagli sensoriali: è un insieme di strutture causali, oggetti persistenti, forze e vincoli fisici, intenzioni e obiettivi altrui — tutti fenomeni che si manifestano solo a livelli di astrazione molto superiori a quello del pixel.

È importante sottolineare che la critica al regime del pixel non equivale a una critica alla ricostruzione come principio epistemico generale. La ricostruzione è un obiettivo potente e legittimo in molti contesti: permette di valutare quantitativamente la qualità di un modello generativo, fornisce un segnale supervisionato denso che non richiede annotazioni umane, e può produrre rappresentazioni utili per compiti di basso livello come la super-risoluzione o la

denoising. Il problema non è nella ricostruzione in quanto tale, ma nella ricostruzione *nello spazio delle osservazioni sensoriali grezze* come obiettivo *primario* dell'apprendimento delle rappresentazioni per sistemi che devono ragionare sul mondo. La distinzione è sottile ma cruciale, e la sezione I.5 la formalizza nella tassonomia degli obiettivi autosupervisionati.

Un'ulteriore considerazione riguarda l'asimmetria fra ricostruzione e predizione. La ricostruzione di una regione mascherata richiede di produrre il segnale sensoriale nella sua interezza — inclusa la texture fine, il rumore di sensore, le variazioni di illuminazione locale — mentre la predizione latente richiede solo di catturare le strutture semantiche che discriminano la regione mascherata da tutte le altre regioni plausibili. Questa asimmetria di richiesta informativa è alla radice del vantaggio strutturale della predizione latente: un modello che deve predire a livello di pixel deve essere più capace, non più intelligente.

I.3 — La questione del collasso delle rappresentazioni

Prima di presentare l'architettura JEPA, è necessario comprendere un ulteriore ostacolo tecnico che qualsiasi approccio all'apprendimento autosupervisionato deve affrontare: il collasso delle rappresentazioni (representation collapse). Il problema si pone ogni volta che si tenta di apprendere rappresentazioni nello spazio latente piuttosto che nello spazio delle osservazioni grezze.

Supponiamo di voler addestrare un encoder $E_\theta: \mathbb{R}^d \to \mathbb{R}^k$ con $k \ll d$ in modo tale che rappresentazioni di input semanticamente simili siano vicine nello spazio latente. Un approccio diretto — addestrare E_θ a minimizzare la distanza tra rappresentazioni di immagini simili — converge inevitabilmente alla soluzione degenere $E_\theta(\mathbf{x}) = \mathbf{c}$ per ogni $\mathbf{x}$ e per qualche costante $\mathbf{c} \in \mathbb{R}^k$. Questa soluzione, detta collasso totale, minimizza trivialmente qualsiasi perdita che penalizza le distanze tra rappresentazioni di coppie positive, poiché tutte le rappresentazioni diventano identiche.

Esistono diverse strategie per prevenire il collasso. La più ovvia è l'apprendimento contrastivo, che introduce esplicitamente coppie negative: si chiede al modello di avvicinare le rappresentazioni di copie aumentate dello stesso campione (coppie positive) e di allontanare le rappresentazioni di campioni distinti (coppie negative). Formalmente, la perdita InfoNCE — una delle formulazioni più diffuse — per un batch di N campioni è:

$$\mathcal{L}_{\text{InfoNCE}}(\theta) = -\,\mathbb{E}\left[\log\left[\exp(z_i \cdot z_i^{+} / \tau) \,/\, \Sigma_{=1}^{N} \exp(z_i \cdot z_j / \tau)\right]\right]$$

dove $z_i = E_\theta(\mathbf{x}_i) \,/\, \|E_\theta(\mathbf{x}_i)\|$ è la rappresentazione normalizzata dell'i-esimo campione, z_i^{+} è la rappresentazione della sua vista aumentata, τ è una temperatura e la somma al denominatore include sia i positivi che i negativi. Questa perdita è una stima del negativo dell'informazione mutua tra le due viste, e la sua massimizzazione spinge il modello a preservare le informazioni condivise tra le viste mantenendo separati i campioni distinti.

Tuttavia l'apprendimento contrastivo ha i propri difetti. Richiede batch di grandi dimensioni per avere una buona stima delle distribuzioni negative, introduce una dipendenza dalla scelta delle augmentazioni che codifica ipotesi esplicite su quali variazioni sono irrilevanti — ipotesi che devono essere ingegnerizzate manualmente per ciascun dominio — e non si presta in modo naturale alla predizione di strutture latenti per regioni mascherate, che è ciò che JEPA cerca di realizzare.

Approcci come BYOL (Bootstrap Your Own Latent) e SimSiam hanno dimostrato che il collasso può essere prevenuto anche senza campioni negativi espliciti, a patto di introdurre asimmetrie architetturali tra i due rami della rete: tipicamente un ramo online con backpropagation completa e un ramo target aggiornato con una media esponenziale mobile (EMA) dei pesi del ramo online. Questa asimmetria rompe la simmetria del problema e impedisce il collasso, ma la comprensione teorica del meccanismo preciso rimane incompleta e oggetto di ricerca attiva.

JEPA risolve il problema del collasso in modo diverso, non attraverso l'asimmetria tra viste ma attraverso la struttura predicativa del compito: il modello deve predire la rappresentazione di una regione mascherata partendo dalla rappresentazione del contesto, e la multimodalità residua viene tollerata piuttosto che eliminata. Come si vedrà, l'architettura garantisce la non-degenerazione attraverso un meccanismo più sottile che coinvolge il design del predictor e il suo rapporto con l'encoder target.

I.4 — Epistemologia della rappresentazione: da Piaget a LeCun

È utile, a questo punto, allargare la prospettiva e collocare il problema della rappresentazione nel contesto più ampio della teoria della conoscenza, non per un vezzo culturale ma perché alcune intuizioni fondamentali della psicologia dello sviluppo e della filosofia della mente hanno dirette conseguenze sull'architettura dei sistemi di apprendimento.

Jean Piaget aveva compreso, attraverso decenni di osservazione del comportamento infantile, che la conoscenza del mondo non emerge dall'accumulo passivo di impressioni sensoriali ma dalla costruzione attiva di schemi cognitivi che vengono continuamente testati, falsificati e ristrutturati attraverso l'interazione con l'ambiente. La permanenza dell'oggetto — la comprensione che un oggetto continua ad esistere anche quando non è percepibile — è una conquista cognitiva che il bambino raggiunge verso gli otto mesi attraverso l'interazione motoria, non attraverso l'osservazione passiva. Questo schema piagetiano del conoscere come agire-e-prevedere è straordinariamente consonante con l'architettura dei World Models moderni, come si vedrà nei capitoli successivi.

Più direttamente rilevante è la tradizione dell'Ipotetico-Deduttivismo nella filosofia della scienza, associata a Karl Popper ma con radici in Duhem e Peirce. Secondo questa tradizione, la conoscenza scientifica non si accumula per induzione da dati grezzi ma attraverso la formulazione di ipotesi (modelli del mondo) e la loro messa alla prova attraverso predizioni falsificabili. Un agente intelligente, da questo punto

di vista, è fondamentalmente un costruttore e testatore di modelli, non un catalogatore di correlazioni. L'architettura JEPA può essere letta proprio come un tentativo di costruire reti neurali che apprendono attraverso un processo strutturalmente analogo: non memorizzare l'output sensoriale ma predire rappresentazioni astratte — ipotesi sull'organizzazione del mondo — e aggiornare i parametri sulla base dell'errore predittivo.

Helmholtz, nel XIX secolo, aveva proposto che la percezione sia fondamentalmente un processo di inferenza inconscia (unbewusste Schluss): il cervello non registra passivamente i segnali sensoriali ma li interpreta attivamente attraverso modelli predittivi del mondo, usandoli per generare aspettative e correggendo tali modelli quando le aspettative vengono violate. Questa intuizione è stata formalizzata nel XX secolo da MacKay, successivamente da Rao e Ballard nel loro influente modello predittivo della corteccia visiva (1999), e più recentemente da Karl Friston nel framework della Energia Libera e dell'inferenza attiva. In tutti questi casi, il meccanismo cognitivo fondamentale è la minimizzazione dell'errore predittivo, non la massimizzazione di una funzione di verosimiglianza sull'output sensoriale.

LeCun è consapevole di questa tradizione e ne fa esplicitamente uso. Il suo framework dell'"Autonomous Machine Intelligence" — di cui JEPA è il componente percettivo — è strutturato intorno a moduli che apprendono modelli del mondo, li usano per pianificare e li aggiornano sulla base dell'esperienza. La connessione con la tradizione dell'inferenza predittiva non è ornamentale: ha conseguenze architetturali precise che verranno analizzate nel dettaglio.

La consonanza tra la tradizione epistemologica descritta e l'architettura JEPA non è soltanto analogica: è strutturalmente precisa. Il principio piagetiano dell'assimilazione e dell'accomodazione — con cui gli schemi cognitivi del soggetto vengono prima applicati all'esperienza e poi modificati quando l'esperienza li falsifica — trova un corrispettivo computazionale nel meccanismo EMA del target encoder: il modello del mondo (il target encoder) viene aggiornato lentamente sulla base delle predizioni del contesto encoder, e questo aggiornamento graduale replica il processo di accomodazione cognitiva senza la brusca discontinuità del cambiamento parametrico ad ogni step. Analogamente, la distinzione helmholtziana tra segnale sensoriale e inferenza inconscia si riflette nella separazione tra l'encoder (che costruisce la rappresentazione a partire dall'input sensoriale) e il predittore (che genera ipotesi sulla regione mascherata senza accedervi direttamente).

La tradizione del predictive processing fornisce, in questo senso, non soltanto una fonte di ispirazione ma una sorta di proof of concept biologico: il cervello dei mammiferi implementa da centinaia di milioni di anni una forma di apprendimento attraverso la minimizzazione dell'errore predittivo nello spazio delle rappresentazioni, e lo fa con una efficienza energetica — venti watt per il cervello umano adulto — che nessun sistema artificiale contemporaneo può avvicinare. Questa efficienza non è irrilevante per il progetto AGI: suggerisce che

il paradigma della predizione latente è non soltanto cognitivamente plausibile ma computazionalmente parsimonioso rispetto ai suoi alternativi.

I.5 — Tassonomia degli obiettivi di apprendimento autosupervisionato

Prima di passare alla descrizione formale di JEPA, è utile tracciare una tassonomia degli obiettivi di apprendimento autosupervisionato, poiché JEPA si coloca in modo preciso all'interno di questa tassonomia e la comprensione del suo posto relativo è essenziale per valutarne i pregi e i limiti.

Si possono distinguere quattro classi principali di obiettivi autosupervisionati:

La prima classe comprende gli obiettivi di **ricostruzione nello spazio dell'input**: il modello riceve una versione corrotta o parziale dell'input e deve restituire l'input originale. Esempi includono i denoising autoencoders, i masked autoencoders (MAE di He et al., 2022), i modelli di completamento linguistico BERT e i modelli di diffusione. L'obiettivo è formalizzato come la minimizzazione di una perdita tra l'output del modello e l'input originale nello spazio delle osservazioni. Come discusso sopra, questa classe soffre del problema della multimodalità nelle predizioni e del bias verso caratteristiche sensoriali a scapito di quelle semantiche.

La seconda classe comprende gli obiettivi di **contrastività nello spazio latente**: il modello apprende a costruire rappresentazioni che siano vicine per viste diverse dello stesso campione e lontane per campioni distinti. Esempi includono SimCLR, MoCo, BYOL, SimSiam, BarlowTwins, VICReg. Questi metodi evitano la ricostruzione sensoriale ma richiedono meccanismi aggiuntivi — campioni negativi o asimmetrie architetturali — per prevenire il collasso. La forma dell'obiettivo è tipicamente:

$$\mathcal{L}_{\text{contrast}}(\theta) = \mathcal{L}_{\text{pos}}(\theta) + \lambda \cdot \mathcal{L}_{\text{neg}}(\theta)$$

dove $\mathcal{L}_{\text{pos}}$ penalizza la distanza tra rappresentazioni di coppie positive e $\mathcal{L}_{\text{neg}}$ penalizza la similarità tra coppie negative, con λ che bilancia i due termini.

La terza classe comprende gli obiettivi di **distillazione della conoscenza**: un encoder student apprende a imitare le rappresentazioni prodotte da un encoder teacher, che può essere una versione precedente dello stesso encoder (EMA, come in BYOL) o un encoder preaddestrato su compiti diversi (come in DINO e DINOv2). Questi metodi hanno prodotto rappresentazioni di qualità eccezionale per la classificazione delle immagini e il segmento denso, ma la comprensione teorica del perché funzionino rimane incompleta.

La quarta classe — quella in cui si colloca JEPA — comprende gli obiettivi di **predizione nello spazio latente**: il modello deve predire la rappresentazione latente di una parte dell'input partendo dalla rappresentazione latente di un'altra parte. L'elemento distintivo è che sia l'input che il target della predizione sono nello spazio latente, non nello spazio delle osservazioni. Questa classe risolve il problema della multimodalità sensoriale perché le rappresentazioni latenti sono (si

spera) invarianti rispetto alle varianti irrilevanti: due immagini che differiscono solo nel rumore di sensore o nella precisa tonalità di un pixel dovrebbero avere rappresentazioni latenti quasi identiche, eliminando così la falsa multimodalità che affligge la predizione a livello di pixel.

Il prezzo da pagare è la necessità di garantire la non-degenerazione delle rappresentazioni attraverso meccanismi diversi da quelli della ricostruzione (dove la degenerazione è naturalmente prevenuta dalla fedeltà all'input) o dalla contrastività (dove è prevenuta dai campioni negativi). Come si vedrà, JEPA risolve questo problema attraverso la combinazione di un predictor espressivo ma separato dall'encoder e di un encoder target aggiornato con EMA — combinazione che crea un delicato equilibrio dinamico tra la tendenza del sistema a collassare verso rappresentazioni costanti e la necessità di mantenere la variabilità sufficiente a rendere il compito predittivo non banale.

I.6 — Il ruolo della struttura del mondo nella progettazione delle architetture

Una delle idee più feconde — e meno discusse nella letteratura mainstream sull'apprendimento automatico — è che l'architettura di un sistema di apprendimento dovrebbe riflettere la struttura del dominio che il sistema cerca di modellare. Questa idea, variamente espressa come "inductive bias", "structure inductive bias" o "architectural prior", ha radici in decenni di lavoro in statistica bayesiana, teorie dell'apprendimento basate sul rasoio di Occam e studi sulla generalizzazione out-of-distribution.

Nel contesto specifico del percepire e del ragionare sul mondo fisico, la struttura rilevante è ricca e ben caratterizzata: il mondo è composto di oggetti che persistono nel tempo, che si muovono in modo localmente continuo, che interagiscono seguendo principi fisici approssimativamente invarianti rispetto al tempo e alla posizione. Gli agenti — siano essi umani, animali o robot — devono pianificare sequenze di azioni su orizzonti temporali variabili, il che richiede la capacità di simulare mentalmente le conseguenze delle proprie azioni prima di eseguirle. Questa simulazione mentale è precisamente ciò che un World Model rende possibile.

La connessione tra struttura del mondo e architettura non è soltanto un principio astratto: ha manifestazioni concrete nelle scelte di design che distinguono sistemi capaci da sistemi fragili. Consideriamo tre proprietà strutturali del mondo fisico che ogni architettura AGI deve riflettere.

La persistenza degli oggetti nel tempo. Un oggetto che si sposta dietro uno schermo continua ad esistere con le stesse proprietà fisiche. Questo fatto — ovvio per un bambino di otto mesi — non è catturato da nessun sistema che apprende rappresentazioni puramente locali nel tempo. Un modello che elabora frame video indipendentemente non può ragionare sulla persistenza; un modello con stato latente ricorrente può approssimarla. La scelta di includere stati ricorrenti nel

World Model non è quindi una preferenza tecnica ma la riflessione architetturale di una proprietà fondamentale del dominio.

La struttura causale delle interazioni fisiche. Quando un oggetto A spinge un oggetto B, il movimento di B è causato da A — non correlato con esso. La distinzione tra causazione e correlazione è cruciale per la pianificazione: un agente che confonde le due non può predire le conseguenze delle proprie azioni su oggetti mai toccati prima. Costruire questa distinzione in un'architettura richiede meccanismi di intervento espliciti — la capacità di simulare il mondo condizionalmente a specifiche azioni dell'agente — che i soli modelli predittivi passivi non possono fornire.

La composizionalità delle configurazioni fisiche. Una scena complessa è composta di oggetti semplici in relazioni spaziali specifiche. Un sistema che rappresenta la scena come un unico vettore denso non può facilmente "decomporre" la scena nei suoi costituenti, ragionare su di essi separatamente, e poi "ricomporre" le predizioni. Le architetture basate su slot — dove ogni slot rappresenta un oggetto distinto — incorporano questa proprietà composizionale in modo strutturale; le architetture basate su patch o token la acquisiscono solo implicitamente, e con minore robustezza fuori distribuzione.

Sia M_θ un modello del mondo parametrizzato da θ.

Sia M_θ un modello del mondo parametrizzato da θ. Il compito fondamentale di M_θ è predire lo stato futuro del mondo z_{t+1} a partire dallo stato presente z_t e dall'azione a_t:

$z_{t+1} \approx M_\theta(z_t, a_t)$

oppure, in versione stocastica, campionare da una distribuzione:

$z_{t+1} \sim p_\theta(z_{t+1} \mid z_t, a_t)$

dove z_t appartiene a uno spazio latente $\mathcal{Z}$ che non è necessariamente lo spazio delle osservazioni sensoriali. L'obiettivo dell'apprendimento è trovare i parametri θ tali che M_θ sia una buona approssimazione della dinamica reale del mondo per le distribuzioni di stati e azioni rilevanti per l'agente.

La connessione con JEPA è immediata: JEPA fornisce il meccanismo per imparare lo spazio $\mathcal{Z}$ stesso, ovvero per costruire rappresentazioni latenti del mondo che siano sufficientemente astratte da sopprimere le varianti sensoriali irrilevanti, sufficientemente informative da supportare predizioni utili, e sufficientemente strutturate da permettere la composizione e la generalizzazione. JEPA e i World Models non sono dunque architetture concorrenti ma moduli complementari: JEPA è il percettore che costruisce $\mathcal{Z}$, il World Model è il predittore che ragiona su $\mathcal{Z}$.

Questa divisione del lavoro ha una precisa corrispondenza con la struttura del cervello dei mammiferi, dove le cortecce sensoriali (visiva, uditiva,

somatosensoriale) costruiscono rappresentazioni gerarchiche degli input sensoriali, mentre le aree frontali e prefrontali usano queste rappresentazioni per pianificare, ragionare e prendere decisioni. Non si intende affermare che le architetture artificiali debbano o possano replicare in dettaglio questa struttura biologica — tale tesi sarebbe difficile da sostenere in assenza di una teoria computazionale completa del cervello — ma che la convergenza funzionale suggerisce principi di progettazione profondi che trascendono le peculiarità implementative.

I.7.A — Perché la matematica è inseparabile dalla comprensione architetturale

Una nota metodologica sulla struttura della trattazione che segue. Le derivazioni matematiche — dalla perdita di I-JEPA alla teoria PAC-Bayes, dall'analisi spettrale dell'attenzione alle ODE delle traiettorie latenti — non vengono presentate per dimostrare rigore formale né come esercizi tecnici separabili dalla narrativa. Vengono presentate perché in questo campo la matematica non è il "linguaggio" con cui si esprimono intuizioni già possedute: è lo strumento attraverso il quale le intuizioni diventano precise abbastanza da essere verificate, falsificate o generalizzate.

Un esempio illuminante: l'intuizione che "la predizione latente produce rappresentazioni migliori della ricostruzione sensoriale" è vera ma vaga. Diventa precisa — e verificabile — soltanto quando si può derivare che la distribuzione ottimale degli embedding per qualsiasi compito di probing è gaussiana isotrópica (Teorema 1 di LeJEPA), che questa distribuzione è raggiungibile attraverso SIGReg con complessità lineare (Teorema 4), e che la loss di training con questa regolarizzazione correla con la performance downstream con un coefficiente di Spearman superiore al 94% (risultato empirico). Senza la matematica, si ha un'impressione; con la matematica, si ha una comprensione strutturale.

Il lettore non matematicamente inclinato può in prima lettura seguire la narrativa e tornare alle derivazioni successivamente: la struttura del testo permette entrambi i percorsi. Ma chi legge la matematica troverà spesso che essa rivela connessioni non ovvie — come la connessione tra il meccanismo EMA e il principio variazionale, o tra la Geodesica Hypothesis e il Principio di Minima Azione — che arricchiscono la comprensione dell'architettura molto oltre quanto possibile con la sola narrazione.

I.7 — Struttura della monografia

Nei capitoli che seguono si costruisce una trattazione progressiva e formalmente rigorosa che parte dalle basi matematiche dell'apprendimento delle rappresentazioni (Capitolo II), procede alla descrizione dettagliata dell'architettura JEPA nella sua variante originale e nelle sue estensioni (Capitolo III), sviluppa poi la storia e il formalismo dei World Models dalle prime formulazioni di Schmidhuber ai modelli di stato latente recurrente di Hafner (Capitolo IV), esplora il legame sintetico tra le due famiglie di architetture nella visione di LeCun

(Capitolo V), affronta infine le questioni di fattibilità computazionale, sostenibilità energetica e scala verso l'AGI (Capitoli VI e VII), per concludere con riflessioni di carattere filosofico sul significato di "comprensione" e "intelligenza" nel contesto di queste architetture (Capitolo VIII).

Il lettore può procedere linearmente oppure affrontare i capitoli tecnici centrali (II–IV) in qualsiasi ordine dopo aver assimilato i fondamenti concettuali del presente capitolo. Le derivazioni matematiche sono presentate nel corpo del testo, non in appendice, perché si ritiene che la comprensione profonda dell'architettura JEPA e dei World Models sia inseparabile dalla comprensione delle strutture matematiche che li definiscono.

Capitolo II — Fondamenti Matematici dell'Apprendimento delle Rappresentazioni

II.1 — Spazi delle rappresentazioni: geometria e topologia

Il punto di partenza rigoroso di qualsiasi analisi dell'apprendimento delle rappresentazioni è la caratterizzazione matematica degli spazi in cui queste rappresentazioni vivono. Sia $\mathcal{X}$ lo spazio delle osservazioni, che nella pratica dell'apprendimento visivo è tipicamente uno spazio di immagini a dimensione finita $\mathbb{R}^{(H\times W\times C)}$ dove H, W sono la risoluzione e C il numero di canali cromatici. Questo spazio ha dimensione nominale che nell'uso moderno supera il milione, ma la distribuzione reale delle immagini naturali occupa una varietà (manifold) di dimensione molto inferiore, stima approssimativa dell'ordine di 10^2–10^3 dimensioni effettive. L'idea fondante dell'apprendimento delle rappresentazioni è che esista una funzione $\varphi_\theta: \mathcal{X} \to \mathcal{Z}$ dove $\mathcal{Z} = \mathbb{R}^k$ con $k \ll \dim(\mathcal{X})$, tale che la varietà delle osservazioni venga proiettata su una varietà di $\mathcal{Z}$ con struttura metricamente più favorevole.

Formalmente, si dice che φ_θ è una buona rappresentazione se soddisfa le seguenti proprietà che si elencano in senso crescente di esigenza:

La prima proprietà è la **separabilità semantica**: per ogni coppia di classi semantiche $C_1, C_2 \subset \mathcal{X}$ con $C_1 \cap C_2 = \varnothing$ nel senso dell'annotazione umana, esiste una distanza $d: \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ tale che:

$$\mathbb{E}_{x\in C_1,\, x'\in C_2}\,[d(\varphi_\theta(x), \varphi_\theta(x'))] \gg \mathbb{E}_{x,x'\in C_1}\,[d(\varphi_\theta(x), \varphi_\theta(x'))]$$

vale a dire che la distanza inter-classe è molto maggiore della distanza intra-classe.

La seconda proprietà è la **invarianza alle trasformazioni irrilevanti**: sia $\mathcal{T}$ un gruppo di trasformazioni — rotazioni, traslazioni, variazioni di illuminazione, trasformazioni affini del colore — che non modificano il contenuto semantico dell'immagine. Si chiede che:

$$\forall\, t \in \mathcal{T},\ \forall\, x \in \mathcal{X}:\ \varphi_\theta(t(x)) \approx \varphi_\theta(x)$$

dove l'approssimazione è nel senso di una distanza piccola in $\mathcal{Z}$.

La terza proprietà, più sofisticata, è la **equivarianza alle trasformazioni rilevanti**: sia $\mathcal{S}$ un gruppo di trasformazioni semanticamente significative — cambiamento di punto di vista, illuminazione direzionale, postura di un corpo articolato. Si chiede che esista una rappresentazione $\varrho: \mathcal{S} \to \mathrm{Aut}(\mathcal{Z})$ del gruppo in automorfismi di $\mathcal{Z}$ tale che:

$$\varphi_\theta(s(x)) = \varrho(s)(\varphi_\theta(x))$$

ovvero che la trasformazione si rifletta in modo strutturato nello spazio latente. Un sistema che possiede equivarianza può "capire" come il proprio stato interno cambierà in seguito a trasformazioni note, il che è essenziale per la pianificazione.

La quarta proprietà è la **disentanglement**: le dimensioni dello spazio latente corrispondono a fattori di variazione indipendenti del mondo. Formalmente, sia $z = \varphi_\theta(x) = (z_1, z_2, ..., z_k)$. Si chiede che i z_i siano statisticamente indipendenti sotto la distribuzione dei dati e che ciascuno z_i corrisponda a un fattore causale identificabile del mondo (forma, colore, posizione, identità, etc.). Il disentanglement è una proprietà molto più forte delle precedenti e non è in generale garantita dall'apprendimento autosupervisionato standard senza vincoli architetturali o regolarizzazioni esplicite. La sua realizzazione completa rimane un problema aperto.

II.2 — Teoria dell'informazione e obiettivi di apprendimento

Il collegamento tra la teoria dell'informazione di Shannon e l'apprendimento delle rappresentazioni è profondo e non meramente formale. L'informazione mutua $I(X; Z)$ tra la variabile aleatoria di input X e la rappresentazione $Z = \varphi_\theta(X)$ è una misura naturale della quantità di informazione sull'input che viene preservata dalla rappresentazione:

$$I(X; Z) = H(X) - H(X \mid Z) = H(Z) - H(Z \mid X)$$

dove $H(\cdot)$ denota l'entropia di Shannon e $H(\cdot \mid \cdot)$ l'entropia condizionale. Una rappresentazione trivialmente informativa è quella identità $\varphi_\theta = \mathrm{id}$, che ha $I(X; Z) = H(X)$ ma che ovviamente non è di alcuna utilità per la generalizzazione. Il principio dell'Information Bottleneck (Tishby, Pereira, Bialek, 1999) suggerisce che una buona rappresentazione massimizza $I(Z; Y)$ — l'informazione su un target di previsione Y — minimizzando al contempo $I(X; Z)$ — la quantità di informazione sull'input che viene trattenuta:

$$\min_{\varphi_\theta} \; [\, -I(Z; Y) + \beta \cdot I(X; Z) \,]$$

dove β è un parametro che bilancia il trade-off tra compressione e predittività. La soluzione ottimale di questo problema di ottimizzazione produce rappresentazioni sufficientemente compresse da filtrare il rumore e le varianti irrilevanti, ma sufficientemente informative da supportare la predizione di Y.

Nel contesto autosupervisionato, dove Y non è un'etichetta esterna ma una parte dello stesso input (ad esempio la porzione mascherata), il problema dell'Information Bottleneck assume una forma particolare. In I-JEPA, il target Y è la rappresentazione latente della regione mascherata, e la minimizzazione di I(X_C; Z_C) con Z_C = E_θ(X_C) — dove E_θ è l'encoder del contesto — è implicitamente incoraggiata dall'uso di grandi regioni di mascheramento che privano il contesto di informazioni locali dettagliate, costringendo l'encoder a catturare strutture globali.

Un'analisi più fine rivela che l'obiettivo di I-JEPA può essere scritto come:

$$\mathcal{L}_IJEPA(\theta, \varphi) = \mathbb{E}_{X, M} [\| s_\varphi(E_\theta(X_C), M) - sg(\bar{E}_\theta(X_T)) \|^2]$$

dove E_θ è l'encoder del contesto con parametri θ, $\bar{E}_\theta$ è il target encoder con parametri $\bar{\theta}$ aggiornati via EMA da θ, s_φ è il predictor con parametri φ, M è la maschera che indica quali regioni sono il target, e sg(·) denota l'operazione stop-gradient che impedisce al gradiente di fluire attraverso il target encoder. La derivazione formale di questa perdita come caso particolare di un obiettivo di Information Bottleneck latente richiede alcune assunzioni sulla struttura delle distribuzioni condizionali, ma l'intuizione di fondo è solida.

II.3 — Il formalismo variazionale e il suo ruolo

Per comprendere perché I-JEPA utilizzi un target encoder EMA piuttosto che minimizzare direttamente la distanza tra le rappresentazioni dei due encoder identici, è necessario analizzare il problema dell'apprendimento in termini variazionali.

Il formalismo variazionale ha una storia ricca che precede l'apprendimento automatico moderno di almeno due secoli. Il calcolo delle variazioni — sviluppato da Euler, Lagrange e poi da Hamilton nel XVIII e XIX secolo — si occupa di trovare funzioni che rendono stazionari certi funzionali, ovvero integrali che dipendono non da numeri ma da funzioni. La connessione con l'apprendimento delle rappresentazioni è profonda: trovare l'encoder ottimale E_θ* è un problema variazionale in cui il funzionale da ottimizzare è la perdita di predizione sul dominio dei parametri θ.

Questa connessione non è puramente formale. Il Principio di Minima Azione della fisica classica — che afferma che le traiettorie fisicamente realizzate sono quelle che rendono stazionaria l'azione $S = \int L(q, \dot{q}, t)\, dt$, dove L è la Lagrangiana del sistema — ha un analogo nell'apprendimento automatico: le rappresentazioni apprese dalla discesa del gradiente su un corpus di dati tendono a concentrarsi nelle regioni dello spazio dei parametri dove la curvatura della perdita è bassa, che corrispondono a soluzioni "stabili" nel senso variazionale. La regolarizzazione implicita della discesa del gradiente stocastica — il fatto che SGD non converge a qualsiasi minimo ma tende verso minimi con certi caratteristiche geometriche — è un'istanza di questo principio.

Nel contesto specifico di I-JEPA, il problema variazionale è particolarmente ricco perché coinvolge due funzioni da ottimizzare simultaneamente: l'encoder E_θ e il predittore P_φ. Non si tratta di un semplice problema di ottimizzazione congiunta (che convergerebbe alla soluzione collassata come si vedrà) ma di un problema di equilibrio — trovare una coppia (E_θ, P_φ) che sia stabile rispetto alle perturbazioni locali, nel senso che piccoli cambiamenti nell'encoder inducono cambiamenti nel predittore che tendono a restaurare l'equilibrio. Il meccanismo EMA trasforma questo problema di equilibrio in un problema di inseguimento controllato: il predittore insegue continuamente il target encoder, che si muove lentamente abbastanza da permettere all'inseguimento di essere stabile.

Consideriamo il problema di apprendere un encoder $E_\theta: \mathcal{X} \to \mathcal{Z}$

Consideriamo il problema di apprendere un encoder $E_\theta: \mathcal{X} \to \mathcal{Z}$ e un predictor $P_\varphi: \mathcal{Z} \times \mathcal{K} \to \mathcal{Z}$, dove $\mathcal{K}$ è lo spazio delle maschere, tali che $P_\varphi(E_\theta(x_C), m) \approx E_\theta(x_T)$ per ogni coppia (x_C, x_T) di contesto e target ottenuta mascherando l'input x con maschera m. Se si usa lo stesso encoder E_θ per produrre sia il contesto che il target, e si minimizza la perdita:

$$\mathcal{L}_{naive}(\theta, \varphi) = \mathbb{E}\left[\| P_\varphi(E_\theta(x_C), m) - E_\theta(x_T) \|^2 \right]$$

il sistema ha a disposizione una soluzione degenere: $E_\theta(x) = \mathbf{0}$ per ogni x, $P_\varphi(z, m) = \mathbf{0}$ per ogni z e m. Con questa scelta, la perdita è identicamente zero per ogni campione, ma le rappresentazioni non contengono alcuna informazione utile. Questo è il collasso totale.

La presenza del predictor P_φ, per quanto sorprendente a prima vista, non impedisce di per sé il collasso: anche con un predictor espressivo il sistema può convergere alla soluzione nulla. Ciò che impedisce il collasso in I-JEPA è la combinazione di due fattori. Il primo è l'asimmetria tra il contesto encoder (con gradiente) e il target encoder (senza gradiente, aggiornato con EMA): poiché i parametri del target encoder sono una versione temporalmente ritardata e mediata del contesto encoder, non è possibile sincronizzare i due encoder verso la soluzione collassata in modo troppo rapido. Il secondo è la natura del predictor: poiché il predictor deve mappare rappresentazioni del contesto a rappresentazioni del target usando solo la maschera come informazione aggiuntiva, esso deve risolvere un problema non banale di "completamento" nello spazio latente, il che crea una pressione verso la conservazione dell'informazione semantica nelle rappresentazioni.

Più formalmente, si può mostrare che la dinamica di aggiornamento EMA:

$$\vartheta \leftarrow \tau \cdot \vartheta + (1 - \tau) \cdot \theta, \tau \in [0.996, 0.999]$$

introduce un ritardo temporale che rende impossibile al contesto encoder di "inseguire" il target encoder verso la soluzione collassata abbastanza velocemente da convergere ad essa. La velocità di convergenza verso il collasso totale è proporzionale a $(1 - \tau)$ / (learning rate), e con τ sufficientemente prossimo a 1,

questa velocità è così bassa che il sistema trova invece soluzioni non degenerate che minimizzano la perdita predittiva.

Il parametro τ è dunque un iperparametro critico che bilancia la stabilità del training (valori alti di τ prevengono il collasso ma possono rallentare l'apprendimento) e la qualità delle rappresentazioni (valori troppo alti rendono il target encoder troppo stabile e quindi il compito predittivo poco informativo).

II.4 — Teoria delle funzioni di base e proiezioni su spazi di Hilbert

Un'altra lente utile per comprendere l'apprendimento delle rappresentazioni è quella della teoria delle funzioni di base negli spazi di Hilbert a nucleo riproducente (RKHS). Un encoder $\varphi_\theta: \mathcal{X} \to \mathbb{R}^k$ può essere visto come una proiezione dell'input su uno spazio di feature finito-dimensionale, dove ciascuna delle k componenti di $\varphi_\theta(x)$ è una funzione base applicata all'input. Dal punto di vista dell'approssimazione funzionale, la questione centrale è: quale famiglia di funzioni è necessaria per rappresentare adeguatamente le quantità del mondo che un agente deve predire?

Per le immagini naturali, la risposta classica — derivata dalla tradizione delle wavelet e della teoria delle strutture multiscala — è che le feature visivamente rilevanti sono localizzate in frequenza e in spazio, e che una buona base di rappresentazione dovrebbe essere gerarchica: feature di basso livello (bordi, texture) nelle prime rappresentazioni, feature di alto livello (parti di oggetti, configurazioni semantiche) nelle rappresentazioni più profonde. Le reti convoluzionali profonde imparano approssimativamente questa struttura gerarchica, e i Vision Transformers la acquisiscono attraverso l'attenzione multi-testa che permette di modellare dipendenze a lungo raggio fin dai livelli più bassi.

L'architettura JEPA utilizza tipicamente un Vision Transformer (ViT) come encoder. In un ViT, l'input viene suddiviso in N patch non sovrapposte di dimensione P×P pixel ciascuna, che vengono proiettate linearmente in vettori di dimensione d:

$e_i = W_E \cdot p_i + pos_i, i = 1, ..., N$

dove $p_i \in \mathbb{R}^(P^2C)$ è il vettore di pixel della i-esima patch, $W_E \in \mathbb{R}^(d \times P^2C)$ è la matrice di proiezione apprendibile, e pos_i è l'embedding posizionale. La sequenza di vettori $(e_1, ..., e_N)$ viene quindi processata da L blocchi Transformer, ciascuno dei quali applica self-attention multi-testa (MHSA) seguita da un feed-forward network (FFN):

$MHSA(H) = Concat(head_1, ..., head_h) \cdot W_O$

dove:

$head_j = Attention(H W_j^Q, H W_j^K, H W_j^V)$

Attention(Q, K, V) = softmax(Q K^T / √(d_k)) · V

con W_j^Q, W_j^K, W_j^V ∈ ℝ^(d × d_k) le matrici di proiezione per query, chiave e valore dell'j-esima testa, d_k = d/h la dimensione di ciascuna testa, e W_O ∈ ℝ^(d × d) la proiezione di output. Il FFN è tipicamente:

FFN(x) = max(0, x W_1 + b_1) W_2 + b_2

oppure, nelle versioni più recenti, usa l'attivazione GELU:

FFN(x) = GELU(x W_1 + b_1) W_2 + b_2

con W_1 ∈ ℝ^(d × 4d), W_2 ∈ ℝ^(4d × d).

Ogni blocco Transformer applica layer normalization prima dell'attenzione e del FFN (Pre-LN normalization):

H' = H + MHSA(LN(H))

H'' = H' + FFN(LN(H'))

La rappresentazione finale dell'intera immagine può essere ottenuta attraverso un token di classificazione [CLS] aggiunto alla sequenza, oppure come media pooling sulle rappresentazioni delle patch al layer finale. ### **II.4.A — Disentanglement, indipendenza statistica e la difficoltà del problema**

Il disentanglement — la quarta proprietà nella scala crescente di esigenza introdotta nella sezione II.1 — merita una trattazione approfondita perché rappresenta il confine tra ciò che i metodi autosupervisionati moderni possono e non possono garantire. Informalmente, una rappresentazione è disentangled se le sue dimensioni corrispondono a fattori di variazione indipendenti del mondo: una dimensione controlla la forma dell'oggetto, un'altra il colore, una terza la posizione, e così via.

La formalizzazione matematica più diffusa del disentanglement richiede indipendenza statistica tra le componenti della rappresentazione. Se $z = (z_1, z_2, \ldots, z_k)$ è il vettore di rappresentazione latente e i fattori generativi del mondo sono $(v_1, v_2, \ldots, v_m)$, si chiede che esista una corrispondenza biunivoca tra i z_i e i v_j tale che $p(z_1, \ldots, z_k) = \prod_i p(z_i)$ *(indipendenza marginale) e che z_i sia funzione di* $v\{f(i)\}$ e non di altri fattori generativi.

Il teorema di impossibilità di Locatello et al. (2019) ha dimostrato un risultato sorprendente: senza inductive bias espliciti o supervisione, il disentanglement non è identificabile a partire dai dati. Formalmente, per qualsiasi distribuzione generativa con fattori latenti indipendenti e per qualsiasi metodo di apprendimento del disentanglement basato esclusivamente sulla massimizzazione dell'ELBO (come i β-VAE), esistono infinite permutazioni e riscalature dello spazio latente che producono rappresentazioni ugualmente verosimili con strutture di disentanglement diverse. La scelta tra queste rappresentazioni non è determinata dai dati ma dall'inductive bias dell'architettura.

Questo teorema di impossibilità ha implicazioni dirette per JEPA: I-JEPA non può garantire il disentanglement senza vincoli aggiuntivi. Le sue rappresentazioni catturano strutture semantiche globali — separano categorie, preservano relazioni strutturali — ma le dimensioni latenti non corrispondono necessariamente a fattori causali identificabili del mondo. Questa non è una critica a I-JEPA ma la delineazione onesta di ciò che può e non può garantire. Il disentanglement genuino richiederebbe vincoli architetturali aggiuntivi — come l'uso di prior strutturati, di supervione debole sulle trasformazioni, o di architetture slot-based — che vanno oltre la predizione latente standard.

dell'attenzione e sue implicazioni per la rappresentazione**

Uno dei contributi teorici più interessanti degli ultimi anni è l'analisi dell'attenzione del Vision Transformer dal punto di vista spettrale. Dong et al. (2021) hanno mostrato che l'attenzione ripetuta tende a produrre il fenomeno della "over-smoothing": dopo molti layer, le rappresentazioni delle patch tendono a convergere verso un valore comune, perdendo la diversità che rende le rappresentazioni discriminative. Questo fenomeno è l'analogo nella famiglia dei Transformer del problema del "oversmoothing" nei GNN (Graph Neural Networks) ed è causato dalla struttura diffusiva dell'operatore di attenzione.

Formalmente, si può scrivere la matrice di attenzione A = softmax(QK^T / √d_k) come un operatore di aggregazione su un grafo completo pesato. L'i-esima riga di A, A_i = (a_{i1}, ..., a_{iN}) con Σ a_{ij} = 1 e a_{ij} ≥ 0, determina come la rappresentazione del token i viene aggiornata come combinazione lineare convessa delle rappresentazioni di tutti i token. Se A è vicina alla matrice uniforme 1/N · **1****1**^T — il che avviene quando i prodotti scalari Q K^T sono quasi uniformi — allora l'aggregazione produce quasi il valor medio delle rappresentazioni, annullando le differenze intra-patch.

L'analisi spettrale di A mostra che i suoi autovalori λ_1 ≥ λ_2 ≥ ... ≥ λ_N soddisfano 0 ≤ λ_i ≤ 1 e λ_1 = 1 (corrispondente all'autovettore uniforme **1**/√N). Il tasso di convergenza verso il valor medio è determinato dal "gap spettrale" 1 − λ_2: quando questo gap è grande, la convergenza è rapida e l'oversmoothing è severo; quando è piccolo, le componenti di alta frequenza del segnale vengono preservate per più layer.

In I-JEPA, il fenomeno dell'oversmoothing ha conseguenze complesse. Da un lato, una certa aggregazione delle rappresentazioni delle patch è desiderabile: il modello dovrebbe produrre rappresentazioni semanticamente coerenti a livello di regione, che richiedono l'aggregazione di informazioni da patch adiacenti e semanticamente correlate. Dall'altro, se l'aggregazione è eccessiva, il predictor si trova a dover ricostruire strutture locali da rappresentazioni troppo uniformi, il che è difficile o impossibile.

I-JEPA gestisce questo trade-off attraverso la scelta dell'architettura del predictor: si usa tipicamente un Transformer di piccola dimensione (narrow Transformer) che elabora solo le rappresentazioni del contesto con l'aggiunta di query

corrispondenti alle posizioni mascherate. Questo predictor è abbastanza espressivo da apprendere come "completare" le strutture semantiche mancanti, ma abbastanza semplice da non essere in grado di memorizzare banalmente l'intero dataset.

II.6 — Bounds sulla generalizzazione: teoria PAC-Bayes e sue applicazioni

La domanda fondamentale che qualsiasi metodo di apprendimento deve rispondere è: perché la minimizzazione dell'errore empirico su un insieme di addestramento finito dovrebbe produrre buone rappresentazioni per dati mai visti? Il framework teorico più potente per rispondere a questa domanda è la teoria PAC-Bayes (Probably Approximately Correct, Bayes), che permette di derivare bound sulla generalizzazione che dipendono sia dalla complessità del modello che dalla qualità delle rappresentazioni apprese.

Un bound PAC-Bayes per un encoder φ_θ e un decoder lineare w ha la forma:

$$\mathcal{L}_{test}(\theta, w) \leq \mathcal{L}_{train}(\theta, w) + \sqrt{\left(\mathrm{KL}(Q \,\|\, P) + \ln(2\sqrt{m} / \delta)\right) / (2m)}$$

dove m è la dimensione del training set, δ è il livello di confidenza, P è una distribuzione a priori sui parametri (prior), Q è la distribuzione a posteriori (posterior), e $\mathrm{KL}(Q \,\|\, P)$ è la divergenza di Kullback-Leibler tra posterior e prior. Questo bound dice essenzialmente che la generalizzazione è buona quando (a) il modello si adatta bene ai dati di training, e (b) la divergenza KL tra la distribuzione dei parametri appresi e quella a priori è piccola — ovvero quando i parametri appresi non sono troppo "sorprendenti" rispetto alle aspettative a priori.

Nel contesto dell'apprendimento delle rappresentazioni, questo bound ha una conseguenza importante: i metodi che apprendono rappresentazioni compatte — dove compatto significa approssimativamente che la distribuzione dei pesi è concentrata vicino a un prior ragionevole — dovrebbero generalizzare meglio. Questo fornisce una giustificazione teorica per l'uso di regolarizzazioni come il weight decay, il dropout, e altri meccanismi di compressione implicita.

Per I-JEPA in particolare, si può argomentare che l'obiettivo di predire rappresentazioni latenti piuttosto che pixel grezzi produce encoder intrinsecamente più compatti nel senso della teoria PAC-Bayes. La ragione intuitiva è che le rappresentazioni latenti di alta semanticità sono più stabili rispetto alle variazioni dell'input — ovvero $\varphi_\theta(x) \approx \varphi_\theta(x + \varepsilon)$ per piccole perturbazioni ε — il che corrisponde a una minore Lipschitz constant della funzione φ_θ, che a sua volta è associata a un minor norm dei pesi. Formalizzare questa intuizione richiede assunzioni sulla struttura della distribuzione dei dati che vanno oltre gli scopi del presente capitolo, ma la connessione è teoricamente motivata.

II.7 — Geometria dell'energia libera e analisi della loss landscape

Un'ulteriore prospettiva teorica di grande rilevanza è quella della geometria della funzione di perdita in relazione allo spazio dei parametri. La minimizzazione della

perdita nell'apprendimento profondo avviene attraverso varianti della discesa stocastica del gradiente su un landscape ad altissima dimensionalità — tipicamente con 10^7–10^{11} parametri nei modelli moderni — che presenta caratteristiche molto diverse dalle superfici convesse della statistica classica.

Prima di analizzare la geometria specifica della loss landscape di I-JEPA, è utile collocare questa analisi nel contesto più ampio della teoria dell'apprendimento profondo. La comprensione della loss landscape è uno dei problemi più attivi e meno risolti della teoria dell'apprendimento automatico moderno, con implicazioni dirette per la progettazione di algoritmi di ottimizzazione, la scelta degli iperparametri e la previsione del comportamento fuori distribuzione.

La visione classica, ispirata all'ottimizzazione convessa, immaginava la loss landscape come un'unica "scodella" con un minimo globale raggiungibile da qualsiasi punto di partenza. Questa visione è stata progressivamente smontata dalla combinazione di analisi teoriche e osservazioni empiriche. Choromanska et al. (2015) hanno dimostrato, sotto assunzioni semplificatrici sull'architettura della rete e sulla distribuzione dei dati, che i minimi locali della perdita per reti profonde random tendono ad avere valori simili al minimo globale — un risultato noto come l'"equivalenza dei minimi locali". Goodfellow et al. (2015) hanno osservato empiricamente che la perdita varia quasi monotonicamente lungo segmenti lineari che connettono due soluzioni trovate da SGD — il fenomeno del "mode connectivity". Draxler et al. (2018) hanno dimostrato l'esistenza di percorsi quasi-planari che connettono coppie di minimi in molti modelli pratici.

Questi risultati suggeriscono che la loss landscape dei modelli profondi moderni, pur essendo ad altissima dimensionalità e non convessa, ha una struttura molto più regolare di quanto la teoria classica avrebbe previsto. I minimi locali di bassa perdita formano una struttura connessa nello spazio dei parametri — quasi un "continuum di soluzioni" piuttosto che un insieme discreto di ottimi isolati. La SGD naviga questa struttura attraverso un processo che combina discesa del gradiente con esplorazione stocastica, e tende a convergere verso regioni del continuum che hanno proprietà di generalizzazione favorevoli.

Sia $f(\theta)$: $\mathbb{R}^d \to \mathbb{R}$ la funzione di perdita

Sia $f(\theta)$: $\mathbb{R}^d \to \mathbb{R}$ la funzione di perdita, dove $d \gg 1$ è il numero di parametri. La struttura di questa funzione in alta dimensione è dominata da un fenomeno fondamentale: la maggior parte dei punti critici (punti dove $\nabla f(\theta) = 0$) sono punti di sella, non minimi locali. Questo risultato, dimostrato formalmente in alcune versioni semplificate del problema (in particolare per reti neurali con architettura matriciale e sotto assunzioni di distribuzione gaussiana dei pesi) da Dauphin et al. (2014) e successivamente generalizzato, ha conseguenze profonde per la pratica dell'addestramento. In particolare, i minimi locali tendono ad avere valori simili — l'insieme dei minimi locali è approssimativamente connesso attraverso percorsi di bassa perdita nello spazio dei parametri — e la discesa del gradiente stocastica con learning rate appropriato riesce a navigare questo landscape in modo efficace.

Nel caso specifico di I-JEPA, la landscape della perdita ha una struttura addizionale rispetto al caso standard supervisionato: l'esistenza di soluzioni degeneri (collasso delle rappresentazioni) corrisponde a un sottoinsieme di parametri dove la perdita è identicamente zero ma le rappresentazioni non hanno valore. L'obiettivo dell'addestramento è trovare minimi non degeneri — ovvero punti dove la perdita è piccola ma le rappresentazioni sono semanticamente informative.

La presenza del target encoder EMA modifica effettivamente la geometria della loss landscape: poiché il target encoder è aggiornato con ritardo rispetto al contesto encoder, il problema di ottimizzazione è non stazionario nel senso che il target della minimizzazione cambia ad ogni step. Questo introduce effettivamente una pressione verso soluzioni stabili e informative: soluzioni dove le rappresentazioni cambiano lentamente sono quelle che minimizzano la perdita in modo consistente attraverso i cambiamenti del target encoder, e le rappresentazioni semanticamente stabili — ovvero quelle che rispondono solo alle proprietà semantiche dell'input, non ai dettagli sensoriali — sono esattamente quelle che cambiano più lentamente rispetto alle perturbazioni casuali dell'input.

II.8 — Il principio di massima entropia e la struttura della distribuzione latente

Un ultimo principio matematico rilevante per la comprensione di I-JEPA è il principio di massima entropia di Jaynes, che afferma che, in assenza di informazioni aggiuntive, la distribuzione di probabilità da assegnare a una variabile aleatoria dovrebbe essere quella che massimizza l'entropia soggetta ai vincoli imposti dall'informazione disponibile.

Applicato alle rappresentazioni latenti, questo principio suggerisce che le rappresentazioni apprese da un buon encoder dovrebbero avere distribuzione marginale ad alta entropia — ovvero dovrebbero usare tutto lo spazio latente disponibile piuttosto che concentrarsi in regioni ristrette — soggette al vincolo che siano sufficientemente informative sull'input. Questo è esattamente il principio operativo dietro loss come VICReg (Variance-Invariance-Covariance Regularization), che aggiunge esplicitamente penalizzazioni per la varianza troppo bassa e la covarianza troppo alta delle rappresentazioni.

In I-JEPA, non viene usata una regolarizzazione esplicita sulla distribuzione delle rappresentazioni, ma si ottiene implicitamente un risultato simile attraverso la struttura del compito predittivo: poiché il predictor deve produrre rappresentazioni diverse per patch con contenuto semantico diverso, l'encoder viene spinto a produrre rappresentazioni con varianza sufficiente a rendere distinguibili i target. Se le rappresentazioni collassassero verso una distribuzione a bassa entropia, il predictor non potrebbe soddisfare questa richiesta, e la perdita aumenterebbe.

Formalmente, se la distribuzione delle rappresentazioni latenti del target encoder è $p(\bar{z})$ e il predictor produce una distribuzione condizionale $q(\hat{z} \mid z_C, m)$, la perdita di I-JEPA è proporzionale alla divergenza KL media:

$$\mathbb{E}_{z_C, m, \bar{z}} [KL(q(\hat{z} \mid z_C, m) \| \delta(\hat{z} - \bar{z}))]$$

dove δ è la distribuzione di Dirac centrata sul target $\bar{z}$. Minimizzare questa perdita rispetto ai parametri del predictor, con il contesto encoder fisso, è equivalente a fare del predictor il miglior approssimatore possibile della distribuzione condizionale del target. Minimizzare rispetto ai parametri del contesto encoder, con il predictor fisso, spinge l'encoder a produrre rappresentazioni del contesto che siano massimamente informative sul target — ovvero che massimizzino $I(Z_C; \bar{Z}_T)$, l'informazione mutua tra la rappresentazione del contesto e quella del target.

Questa analisi chiude il cerchio con il principio dell'Information Bottleneck presentato nella sezione II.2: I-JEPA implementa, implicitamente e attraverso una struttura architetturale, un apprendimento orientato dalla massimizzazione dell'informazione mutua tra contesto e target nello spazio latente. La differenza rispetto alle formulazioni esplicite dell'Information Bottleneck — come quelle di Tishby o di varianti variazionali successive — è che in I-JEPA questa massimizzazione avviene senza richiedere la specificazione esplicita di una distribuzione a priori sulle rappresentazioni, il che la rende più pratica e scalabile.

Il capitolo seguente costruirà su queste fondamenta per presentare l'architettura I-JEPA nella sua specificità tecnica: le scelte di mascheramento, la struttura del predictor, le varianti multi-block, le procedure di training, e le connessioni empiriche con i benchmark di transfer learning.

I.8 — La violazione empirica delle Scaling Laws come conferma della distinzione tra addestramento tipico e ottimale

Il Capitolo I aveva stabilito la tesi programmatica di questa monografia partendo da una tensione costitutiva: i sistemi di intelligenza artificiale più capaci prodotti dall'umanità presentano simultaneamente capacità straordinarie e limitazioni strutturali che non possono essere superate attraverso la mera accumulazione di parametri, dati e compute. Questa tesi, formulata nel 2022 come critica dei paradigmi generativo-predittivi a livello sensoriale, trovava un supporto indiretto nelle analisi teoriche di LeCun e nei risultati empirici su benchmark di ragionamento — ma mancava di una dimostrazione diretta che la barriera delle Scaling Laws fosse superabile mediante un obiettivo di training qualitativamente diverso, senza modificare l'architettura né aumentare la scala. Semantic Tube Prediction offre precisamente questa dimostrazione.

Il paper di Huang, LeCun e Balestriero (2026) apre con un'affermazione che merita di essere citata nella sua sostanza come cristallizzazione formale della tesi del Capitolo I: le leggi empiriche di scala caratterizzano l'addestramento tipico, non quello ottimale, e il vincolo rigido della legge di potenza è un artefatto degli

obiettivi correnti piuttosto che un limite fisico dell'apprendimento. La dimostrazione è empirica nella forma ma teorica nella sostanza: STP permette a modelli della famiglia Llama-3 (1B, 3B, 8B parametri) di eguagliare l'accuratezza del fine-tuning standard con 1/16 dei dati di addestramento sul dataset NL-RX-SYNTH, con risultati comparabili su GSM8K, Spider, NQ-Open e HellaSwag. Le leggi di Chinchilla predicono, per una riduzione di D di un fattore 16 a parità di N, una crescita della loss proporzionale a (16)^{α_D} con α_D ≈ 0.095, ovvero un deterioramento di (16)^{0.095} ≈ 1.35×. Un modello che raggiunge la medesima accuratezza con D/16 dati viola questa previsione per un fattore di circa 1.35 nella loss, che si traduce in un miglioramento misurabile dell'accuratezza nei task considerati.

Il meccanismo che spiega questa violazione chiarisce la distinzione tra tipico e ottimale a livello fisico. L'obiettivo ℒ_NTP standard ha un rapporto segnale-rumore strutturalmente limitato: minimizzare la cross-entropy rispetto al token successivo non distingue tra le variazioni semanticamente significative degli stati nascosti (componente parallela alla geodetica, segnale) e le variazioni ortogonali (deviazioni dal percorso ottimale, rumore). Il rumore non è un artefatto del dataset ma una conseguenza della discretizzazione: poiché lo stato latente continuo h_t deve essere proiettato in un token discreto x_{t+1} tramite la cella di Voronoi, il training con ℒ_NTP ha solo il requisito che h_t cada nella cella corretta, senza vincoli sulla sua posizione precisa all'interno della cella. Questo lascia un grado di libertà non penalizzato che nel tempo si accumula sotto forma di rumore nella traiettoria latente. ℒ_STP penalizza esattamente questo grado di libertà imponendo che la posizione di h_t all'interno della cella sia allineata con la direzione geodetica, ovvero con il segnale semantico della sequenza.

Il legame con la sezione I.2 sulla critica del regime del pixel è più diretto di quanto appaia a prima vista. La critica del regime del pixel affermava che l'obiettivo di ricostruzione sensoriale dedica capacità rappresentativa alla modellazione di varianze semanticamente irrilevanti — la precisa microtrama di un muro sullo sfondo, il pattern casuale del rumore del sensore — a scapito delle strutture semantiche globali. STP identifica un problema analogo nell'obiettivo autoregressivo per le sequenze: la cross-entropy rispetto al token successivo permette che gli stati nascosti si dispongano arbitrariamente all'interno delle celle di Voronoi dei token corretti, dedicando varianza latente alla codifica di variazioni senza significato semantico. In entrambi i casi, il problema non è nella natura della supervisione (ricostruzione vs predizione del prossimo token) ma nel fatto che la supervisione avviene nello spazio delle osservazioni grezze o della distribuzione discreta dei token, non nello spazio delle strutture semantiche. La soluzione, in entrambi i casi, è un vincolo geometrico sullo spazio latente che forza il modello a organizzare le proprie rappresentazioni secondo la struttura semantica del dominio — la struttura delle patch semanticamente simili in I-JEPA, la struttura geodetica delle traiettorie in STP.

La sezione I.3 aveva discusso il problema del collasso delle rappresentazioni come ostacolo fondamentale all'apprendimento nello spazio latente. LeJEPA risolve

questo problema in modo radicalmente diverso da tutte le soluzioni precedenti: non prevenendo il collasso attraverso euristiche (stop-gradient, asimmetrie architetturali, reti teacher-student) ma rendendolo impossibile per costruzione. La dimostrazione che la gaussiana isotrópica è l'unica minimizzatrice del rischio downstream sotto vincoli scalari sulla covarianza — valida per probing lineare e non lineare — trasforma il problema del collasso da questione pratica a questione formale: una distribuzione collassata è, per definizione, non isotrópica, e quindi non ottimale rispetto a qualsiasi compito downstream. SIGReg forza la distribuzione verso la gaussiana isotrópica non come euristica ma come implementazione di un risultato di ottimalità provato.

Questi due risultati — STP come violazione dimostrabile delle Scaling Laws, LeJEPA come soluzione provata al collasso — costituiscono insieme il più forte sostegno empirico e teorico disponibile alla data di questa monografia per la tesi centrale del Capitolo I: che le limitazioni dei paradigmi dominanti non siano limitazioni della scala computazionale ma limitazioni degli obiettivi di training, e che obiettivi enfatizzati geometricamente possano superarle per ragioni di principio piuttosto che per aggiustamenti tecnici incrementali.

II.9 — Revisione della critica agli approcci contrastivi alla luce della distribuzione ottimale provata

La sezione II.3 della monografia (qui riferita come collocazione nella struttura, anche se non numerata esplicitamente nel documento) aveva discusso il problema delle JEPA rispetto agli approcci contrastivi, notando che i metodi contrastivi come SimCLR, MoCo e DINO prevengono il collasso attraverso l'uso di campioni negativi — coppie di esempi non correlati la cui distanza nello spazio latente viene massimizzata in parallelo alla minimizzazione della distanza delle coppie positive. Il costo computazionale dei campioni negativi scala con la dimensione del batch (quadratico nel caso naive) o richiede strutture di memoria speciali (code in MoCo), e la scelta di quanti e quali campioni negativi usare è un iperparametro delicato.

LeJEPA svela la ragione teorica profonda di questa situazione. Gli approcci contrastivi sono, nella loro essenza, tentativi di imporre alla distribuzione degli embedding una forma isotropica senza derivare esplicitamente qual è la distribuzione ottimale: massimizzare la distanza tra campioni negativi è un modo indiretto di allargare la varianza della distribuzione, che è uno dei due requisiti (l'altro è la zero correlazione tra dimensioni) per la gaussiana isotrópica. Il meccanismo è però incompleto: la massimizzazione delle distanze tra campioni negativi garantisce che la distribuzione non collassi, ma non garantisce che la forma risultante sia isotrópica nel senso preciso richiesto dal Teorema 1 di LeJEPA. In particolare, la massimizzazione delle distanze nei metodi contrastivi tende a produrre distribuzioni uniformi sulla sfera ipersferica (ogni campione è spinto lontano dagli altri), non distribuzioni gaussiane isotrópiche (che hanno densità esponenzialmente decrescente dalla media) — una differenza che diventa

rilevante per compiti downstream che richiedono la struttura radiale della distribuzione, come i metodi kernel che usano funzioni di base gaussiana.

SIGReg fornisce, in questo senso, una soluzione più precisa agli approcci contrastivi: invece di imporre indirettamente (attraverso la separazione dei campioni negativi) una forma distribuzionale imprecisata, impone direttamente (attraverso il confronto delle funzioni caratteristiche proiettate) che la distribuzione sia gaussiana isotrópica nel senso statisticamente rigoroso del Test di Epps-Pulley. La convergenza tra questi due approcci — contrastivo e basato su distribuzione matching — non è quindi superficiale: entrambi cercano di prevenire il collasso distribuendo le rappresentazioni in modo isotropico nello spazio latente, ma LeJEPA è il primo sistema che derivato il bersaglio distributivo esatto da principi di ottimalità e poi implementato un test statistico rigoroso per il raggiungimento di quel bersaglio. Questo trasforma la prevenzione del collasso da ingegneria euristica a ingegneria teoricamente fondata.

Capitolo III — L'Architettura JEPA: Formalismo, Varianti e Analisi Critica

III.1 — Genesi intellettuale e posizionamento nel paesaggio della ricerca

I-JEPA non è un paper empirico nel senso convenzionale — non riporta un nuovo stato dell'arte su un benchmark di classificazione e non propone un'architettura radicalmente nuova rispetto ai predecessori. È qualcosa di più raro nell'apprendimento automatico moderno: un'architettura giustificata, i cui componenti esistono perché senza di essi qualcosa di specifico non funzionerebbe, e le cui scelte di design sono derivate da principi piuttosto che ricavate per ottimizzazione empirica. Comprendere questa struttura giustificativa è il fine di questo capitolo.

La storia intellettuale della predizione latente è più lunga di quanto suggerirebbe la data di pubblicazione di I-JEPA. L'intuizione che le rappresentazioni dovrebbero essere predittive delle une rispetto alle altre — piuttosto che predittive delle osservazioni grezze — attraversa la letteratura sull'apprendimento per rafforzamento come il Successor Representation di Dayan (1993), la tradizione della predictive coding da Rao e Ballard (1999) in poi, e le reti di tipo Contrastive Predictive Coding di Oord et al. (2018). Ciò che distingue JEPA da questi predecessori è la sistematicità con cui la separazione tra encoder e predittore viene progettata per prevenire il collasso, e la chiarezza con cui viene identificato il tipo di struttura semantica che l'encoder dovrebbe apprendere.

Image-JEPA (I-JEPA) viene presentato pubblicamente da LeCun e collaboratori di Meta AI nel giugno 2023 con il paper "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture". L'articolo si inserisce in un dibattito già avanzato sull'apprendimento autosupervisionato delle immagini, ma lo affronta con un'originalità di prospettiva che merita di essere riconosciuta: mentre la

comunità era divisa tra il paradigma contrastivo (SimCLR, MoCo, DINO) e quello ricostruttivo (MAE, BEiT), I-JEPA propone una terza via che non richiede né coppie negative né ricostruzione sensoriale, basandosi invece sulla predizione latente come obiettivo intrinsecamente sufficiente.

È importante sottolineare che l'idea della predizione nello spazio latente non nasce con I-JEPA. Elementi concettuali analoghi si trovano in BYOL (dove un network predice le rappresentazioni del target network a partire da viste aumentate, ma senza mascheramento), in BEiT (dove si predice la tokenizzazione discreta di un VQVAE piuttosto che i pixel grezzi), e in lavori precedenti di LeCun stesso e di altri sulla struttura delle architetture predittive. Il contributo specifico di I-JEPA è la combinazione di tre elementi: il mascheramento semanticamente informativo, il predictor stretto (narrow predictor) che opera solo sulle rappresentazioni del contesto, e l'analisi teorica che mette in relazione queste scelte con la qualità delle rappresentazioni apprese.

III.2 — Descrizione formale completa dell'architettura I-JEPA

Sia $X \in \mathbb{R}^{(H \times W \times 3)}$ un'immagine di input. Si considera una griglia regolare di $N = (H/P) \times (W/P)$ patch di dimensione $P \times P$ pixel, con $P = 16$ nella configurazione standard di ViT-B e ViT-L. L'architettura I-JEPA comprende tre componenti parametrizzati:

Il **context encoder** $E_\theta: \mathbb{R}^{(H \times W \times 3)} \to \mathbb{R}^{(N_C \times d)}$, dove $N_C \leq N$ è il numero di patch del contesto (quelle non mascherate), è un Vision Transformer con parametri θ che trasforma le patch visibili in una sequenza di rappresentazioni latenti di dimensione d. In I-JEPA, a differenza di MAE, il contesto encoder processa solo le patch visibili — non vengono forniti token di maschera come input — il che riduce il costo computazionale e impedisce la "fuga di informazione" dalle posizioni mascherate.

Il **target encoder** $\bar{E}_{\bar{\theta}}: \mathbb{R}^{(H \times W \times 3)} \to \mathbb{R}^{(N \times d)}$, con parametri $\bar{\theta}$ aggiornati tramite EMA, processa l'intera immagine (incluse le patch mascherate) e produce rappresentazioni per tutte le N patch. Le rappresentazioni delle patch mascherate — quelle nel set T di target patch — costituiscono il segnale di supervisione del sistema. Crucialmente, il gradiente non fluisce attraverso $\bar{E}_{\bar{\theta}}$: le sue rappresentazioni sono trattate come target fissi nel corso di un singolo step di aggiornamento.

Il **predictor** $P_\varphi: \mathbb{R}^{(N_C \times d)} \times \mathbb{R}^{(N_T \times d_{pos})} \to \mathbb{R}^{(N_T \times d)}$, con parametri φ, è un Transformer di dimensione ridotta — tipicamente con $d_{pred} < d$ — che riceve in input le rappresentazioni del contesto encoder e un insieme di query corrispondenti alle posizioni delle patch mascherate (codificate solo attraverso l'embedding posizionale, senza informazione di contenuto) e produce rappresentazioni predette per ciascuna posizione mascherata. La dimensione ridotta del predictor è una scelta deliberata, motivata sia da ragioni computazionali che teoriche: un predictor troppo espressivo potrebbe "bypassare" la necessità di

buone rappresentazioni nel contesto encoder imparando trasformazioni molto complesse delle feature del contesto.

L'aggiornamento dei parametri in ciascun passo di addestramento avviene come segue. Per un batch di immagini, si campiona per ciascuna immagine un insieme di target blocks $T = \{T_1, T_2, ..., T_K\}$ — tipicamente $K = 4$ blocchi rettangolari di dimensione variabile — e un contesto block C che copre le patch non incluse in nessun T . Si calcola la perdita:

$$\mathcal{L}(\theta, \varphi) = (1/K) \cdot \Sigma_{=1}^{K} (1/|T|) \cdot \Sigma_{i \in T} \| P_\varphi(E_\theta(X_C), pos_i) - sg(\bar{E}_\vartheta(X)_i) \|^2$$

dove X_C indica l'immagine con le patch dei target bloccate (non fornite all'encoder del contesto), pos_i è l'embedding posizionale della i-esima patch, $\bar{E}_\vartheta(X)_i$ è la rappresentazione della i-esima patch prodotta dal target encoder, e $sg(\cdot)$ è l'operazione stop-gradient. Si calcolano i gradienti $\partial\mathcal{L}/\partial\theta$ e $\partial\mathcal{L}/\partial\varphi$ e si aggiornano θ e φ con un ottimizzatore AdamW. Successivamente si aggiorna il target encoder:

$$\vartheta \leftarrow \tau \cdot \vartheta + (1 - \tau) \cdot \theta$$

con τ che viene incrementato linearmente da $\tau_start = 0.996$ a $\tau_end = 1.0$ nel corso del training, secondo il schedule:

$$\tau(t) = \tau_end - (\tau_end - \tau_start) \cdot \cos(\pi \cdot t / T) / 2$$

dove t è lo step corrente e T il numero totale di step. Questo schedule di EMA è stato introdotto da BYOL e si è dimostrato empiricamente vantaggioso rispetto al valore costante.

III.3 — La strategia di mascheramento: motivazioni teoriche e implementazione

La scelta della strategia di mascheramento è uno degli aspetti più delicati e teoricamente motivati di I-JEPA, e merita un'analisi approfondita. Non è un dettaglio implementativo: determina la natura del compito predittivo e quindi le proprietà delle rappresentazioni che vengono apprese.

Si considerino tre possibili strategie di mascheramento per la selezione dei target blocks:

La prima strategia, che potremmo chiamare **mascheramento casuale uniforme** (come in BERT per il testo o nella versione base di MAE), seleziona le patch target con probabilità uniforme indipendente. Con questa strategia, il modello deve predire patch isolate o piccoli cluster, il che è un compito relativamente facile che non richiede ragionamento semantico a lunga distanza: la patch adiacente nel contesto fornisce informazioni quasi sufficienti per la predizione attraverso interpolazione delle texture.

La seconda strategia, quella effettivamente adottata in I-JEPA, è il **mascheramento a blocchi semanticamente calibrato**. Si campionano blocchi rettangolari contigui di dimensione variabile, con area compresa tra il 15% e il 20% dell'immagine per ciascun blocco target, e il contesto è ottenuto come complementare di tutti i blocchi target con un ulteriore mascheramento aleatorio aggiuntivo. Questa strategia rende il compito predittivo genuinamente difficile: il modello deve predire regioni estese dell'immagine che non possono essere ricostruite per interpolazione locale, ma richiedono comprensione della struttura semantica globale della scena.

La terza strategia possibile — non usata in I-JEPA ma proposta in varianti successive — è il **mascheramento semanticamente guidato**: le regioni da mascherare vengono selezionate in modo da corrispondere a segmenti semantici completi (oggetti, parti di oggetti) identificati da un segmentatore esterno. Questa strategia è la più potente in teoria ma richiede un oracle di segmentazione che non è disponibile in un regime puramente autosupervisionato.

La scelta del mascheramento a blocchi in I-JEPA è motivata dall'analisi seguente. Definiamo la difficoltà intrinseca del compito predittivo per una maschera m come:

$$D(m) = H(Z_T \mid Z_C) - H(Z_T \mid X_T, Z_C)$$

dove $Z_T = \bar{E}_{\bar{\theta}}(X)_T$ è la rappresentazione target, $Z_C = E_\theta(X_C)$ è la rappresentazione del contesto, e X_T è il segnale sensoriale grezzo delle patch target. Il primo termine $H(Z_T \mid Z_C)$ è l'incertezza sulla rappresentazione target dato solo il contesto latente; il secondo termine $H(Z_T \mid X_T, Z_C)$ è l'incertezza residua anche con accesso ai pixel target (questa è l'incertezza intrinseca dovuta alla stocasticità del target encoder). Una difficoltà elevata corrisponde a un compito che richiede informazioni semantiche non locali, il che spinge l'encoder a catturare strutture a lungo raggio.

Con mascheramento uniforme, $D(m)$ è tipicamente basso perché le patch adiacenti nel contesto sono molto informative sulla patch target. Con mascheramento a blocchi grandi, $D(m)$ è tipicamente alto perché la predizione richiede di ragionare sulla struttura globale della scena. I-JEPA calibra la dimensione dei blocchi affinché $D(m)$ sia sufficientemente alto da rendere necessario il ragionamento semantico ma non così alto da rendere il compito impossibile da risolvere nemmeno con rappresentazioni perfette.

III.4 — Il predictor ristretto come principio regolarizzante

Il predictor P_φ in I-JEPA è intenzionalmente piccolo: nella configurazione standard con un context encoder ViT-L (24 layer, $d = 1024$), il predictor ha solo 12 layer e $d_{pred} = 384$, ovvero dimensione latente circa tre volte inferiore. Questa asimmetria ha conseguenze teoriche importanti.

La relazione tra la dimensione del predittore e la qualità delle rappresentazioni apprese dall'encoder è una delle intuizioni più profonde di I-JEPA. Dal punto di vista della teoria dell'informazione, definiamo la capacità del predittore come la quantità massima di informazione mutua che può catturare tra le rappresentazioni del contesto e quelle target: $C_P = \max_{\varphi} I_P(Z_C; \hat{Z}_T)$. Se il predittore ha capacità infinita, può potenzialmente "ricostruire" le rappresentazioni target anche quando il context encoder produce rappresentazioni collassate, semplicemente imparando a mappare la rappresentazione costante del contesto alla rappresentazione target più frequente. Il collasso non viene penalizzato perché il predittore compensa la perdita di informazione nell'encoder.

Al contrario, con un predittore di capacità limitata, il collasso dell'encoder produce una perdita di predizione che non può essere compensata. L'encoder è costretto a mantenere sufficiente variabilità nelle sue rappresentazioni da rendere il compito predittivo risolvibile. Questa variabilità corrisponde all'informazione semantica che vogliamo catturare.

Sia C_P la complessità del predictor, misurata ad esempio dal numero di parametri o dalla sua capacità di approssimazione funzionale nel senso di Vapnik-Chervonenkis. Sia C_E la complessità del contesto encoder. L'invariante chiave è che $C_P \ll C_E$: il predictor non è abbastanza capace da apprendere trasformazioni arbitrarie delle rappresentazioni del contesto. Questo crea una pressione implicita sull'encoder affinché le sue rappresentazioni siano già "allineate" con le rappresentazioni target nel senso che la trasformazione necessaria per mapparle l'una sull'altra sia semplice — ovvero che le rappresentazioni del contesto e del target siano nello stesso spazio semantico, non in spazi arbitrariamente diversi che un predictor complesso potrebbe allineare post-hoc.

Formalmente, si può pensare al predictor come a un vincolo sulla classe di funzioni che il contesto encoder può imparare. Se il predictor può apprendere solo funzioni della forma $P_\varphi(z_C, pos) = W \cdot z_C + b(pos)$ (predictor lineare), allora l'encoder deve produrre rappresentazioni z_C che siano linearmente predittive delle rappresentazioni target. Questo vincolo di linearità è molto forte e tende a produrre spazi latenti con struttura geometrica semplice. Il predictor effettivo di I-JEPA è non-lineare (un Transformer), ma la sua capacità limitata impone comunque una forma debole di vincolo di struttura sulle rappresentazioni.

Questa analisi suggerisce un principio di progettazione generale: nelle architetture predittive latenti, la dimensione del predictor controlla il tipo di struttura che viene richiesta allo spazio latente. Un predictor troppo piccolo produce rappresentazioni con struttura troppo semplice che non catturano le non-linearità semantiche del dominio; un predictor troppo grande lascia troppa flessibilità all'encoder e tende verso rappresentazioni degeneri o difficili da usare per compiti downstream.

III.5 — Analisi della convergenza e proprietà dello spazio latente

Una questione che non ha ancora ricevuto risposta soddisfacente nella letteratura è: a quale distribuzione converge lo spazio latente di I-JEPA dopo il training? Possiamo sviluppare alcune considerazioni teoriche parziali.

La questione della distribuzione a convergenza di I-JEPA è intimamente connessa con il problema dell'unicità delle soluzioni nell'apprendimento non supervisionato. Nel contesto supervisionato, la funzione target è fissata esternamente; nel contesto autosupervisionato, qualsiasi encoder che soddisfa il compito predittivo con errore zero è una soluzione valida, e lo spazio di tali soluzioni può essere enorme.

Introduciamo il concetto di equivalenza predittiva: due encoder E_θ e E_θ' sono predittivamente equivalenti se la perdita di I-JEPA ottimizzata rispetto al predittore è uguale per entrambi. L'insieme delle classi di equivalenza predittive forma lo "spazio delle soluzioni equivalenti" — tipicamente molto grande, includendo tutte le trasformazioni invertibili dello spazio latente che preservano la struttura predittiva. Il meccanismo EMA riduce questo spazio introducendo una simmetria rotta: le trasformazioni che genererebbero soluzioni equivalenti non possono essere applicate senza modificare sia il context che il target encoder in modo coerente, rendendo molte trasformazioni instabili rispetto alla dinamica di aggiornamento.

Sia p*(z) la distribuzione marginale delle rappresentazioni del target encoder sulla distribuzione dei dati naturali. Al termine del training, se il sistema converge a un equilibrio non degenere, questa distribuzione deve soddisfare simultaneamente: (1) deve avere entropia sufficiente da essere informativa sulle strutture semantiche dell'immagine; (2) deve essere tale che il predictor possa approssimare la distribuzione condizionale p*(z_T | z_C, m) in modo ragionevolmente accurato; (3) deve essere stabile rispetto all'aggiornamento EMA, ovvero la distribuzione prodotta dal contesto encoder deve essere "vicina" a quella del target encoder nel senso di una divergenza appropriata.

La condizione (3) è particolarmente interessante. Con l'aggiornamento EMA:

$$\bar{\theta}(t+1) = \tau \cdot \bar{\theta}(t) + (1 - \tau) \cdot \theta(t)$$

la distribuzione delle rappresentazioni del target encoder è una media temporale delle distribuzioni prodotte dal contesto encoder nei passi precedenti:

$$\bar{E}_{\bar{\theta}}(t) \approx (1 - \tau) \cdot \Sigma_{s=0}^{\infty} \tau^s \cdot E_{\theta}(t-s)$$

Questo implica che il target encoder è una versione temporalmente smoothed del contesto encoder, ovvero che le sue rappresentazioni sono meno sensibili alle fluttuazioni dell'ottimizzazione e più stabili rispetto ai trend a lungo termine dell'apprendimento. La distanza tra le distribuzioni del contesto encoder e del target encoder è controllata da τ e dalla velocità di cambiamento di θ nel tempo: con $\tau \approx 0.999$ e learning rate piccolo, le due distribuzioni sono quasi identiche, e il compito predittivo è quasi quello di imparare una "funzione identità latente" con

mascheramento. Con $\tau \approx 0.9$ e learning rate grande, le due distribuzioni possono divergere significativamente, il che rende il compito più difficile ma anche potenzialmente più informativo.

Empiricamente, la configurazione $\tau \in [0.996, 0.999]$ con learning rate iniziale 1.5×10^{-3} decrescente con cosine schedule si è dimostrata la più efficace per ViT-L su ImageNet-1K con batch size 2048. Questo punto dell'hyperparameter space corrisponde a un regime dove le rappresentazioni cambiano abbastanza lentamente da essere stabili ma abbastanza velocemente da rimanere informative.

III.5.2 — La distribuzione ottimale delle rappresentazioni: dalla congettura all'ottimo provato

La domanda che per anni ha aleggiato sulla ricerca nelle architetture a doppio embedding senza mai ricevere risposta soddisfacente è di una semplicità quasi irritante nella sua formulazione: quale distribuzione dovrebbe seguire l'embedding prodotto da un encoder ben addestrato per minimizzare il rischio su compiti downstream arbitrari? Non si trattava soltanto di una questione teorica di eleganza formale, ma di un problema pratico di primaria importanza, poiché la risposta avrebbe dovuto guidare la progettazione dei meccanismi di regolarizzazione impiegati per prevenire il collasso delle rappresentazioni. Invece di rispondere a questa domanda, la comunità aveva proceduto per via empirica — scegliendo stop-gradient, normalizzazioni, architetture asimmetriche teacher-student — accumulando ricette che funzionavano senza che nessuno sapesse con precisione perché. LeJEPA (Balestriero e LeCun, 2025) rompe con questa tradizione ponendosi l'obiettivo opposto: derivare la struttura ottimale dell'embedding da principi di primo ordine, poi costruire l'architettura come conseguenza necessaria di tale derivazione.

La questione viene formalizzata nel modo seguente. Sia $f_\theta : \mathbb{R}^D \to \mathbb{R}^K$ l'encoder con $K \ll D$, e sia $Z = f_\theta(X)$ la variabile aleatoria delle rappresentazioni quando X è campionata dalla distribuzione dei dati. Si chiede: tra tutte le possibili distribuzioni marginali p_Z, soggette a un vincolo scalare sulla covarianza (ad esempio $\mathrm{tr}(\mathrm{Cov}(Z)) = \varkappa$ o $\|\mathrm{Cov}(Z)\|_F = \varkappa$ per qualche costante finita $\varkappa$), quale minimizza il rischio atteso su una famiglia ampia di compiti downstream?

La risposta, dimostrata rigorosamente per il probing lineare e per il probing non-lineare tramite regressione kernel e k-NN radiale, è univoca: la distribuzione ottimale è la gaussiana isotropica, ovvero $p_Z = \mathcal{N}(0, (\varkappa/K) \cdot I_K)$.

Il percorso dimostrativo per il probing lineare si articola in due lemmi distinti, l'uno riguardante la varianza dell'estimatore, l'altro il suo bias. Si consideri il problema dei minimi quadrati con regolarizzazione di Tikhonov, nel quale si stima il parametro $\beta \in \mathbb{R}^K$ come

$$\beta = (Z^\top Z + \lambda_{wd} I)^{-1} Z^\top y,$$

dove $Z \in \mathbb{R}^{N\times K}$ è la matrice degli embedding, $y \in \mathbb{R}^N$ il vettore dei target, e $\lambda_{wd} \geq 0$ il coefficiente di regolarizzazione. La decomposizione spettrale $G = Z^T Z / N = Q \Lambda Q^T$, con Q ortogonale e $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, permette di scrivere il bias dell'estimatore come

$$\mathrm{Bias}(\beta) = \mathbb{E}[\beta] - \beta^* = -\lambda_{wd} (G + \lambda_{wd} I)^{-1} \beta^* = -\lambda_{wd} Q (\Lambda + \lambda_{wd} I)^{-1} Q^T \beta^*.$$

Si valuta ora la norma del bias per due configurazioni di covarianza che condividono la stessa varianza totale, ovvero $\Sigma_k \lambda_k = N \varkappa$ (con N campioni). La configurazione anisotropica ha autovalori distinti $\{\lambda_k\}$, mentre la configurazione isotropica ha tutti gli autovalori uguali a $\varkappa$. Scegliendo come parametro vero $\beta^* = \varkappa_0 \cdot q_K$, dove q_K è l'autovettore corrispondente al più piccolo autovalore λ_K della configurazione anisotropica, si ottiene

$$\|\mathrm{Bias}(\beta)\|_{aniso} = \lambda_{wd} / (\lambda_K + \lambda_{wd}) \cdot \|\beta^*\|,$$

$$\|\mathrm{Bias}(\beta)\|_{iso} = \lambda_{wd} / (\varkappa + \lambda_{wd}) \cdot \|\beta^*\|.$$

Poiché $\lambda_K < \varkappa$ per qualsiasi distribuzione strettamente anisotropica (e questa disuguaglianza è stretta ogni volta che gli autovalori non sono tutti uguali), si ha

$$\lambda_{wd} / (\lambda_K + \lambda_{wd}) > \lambda_{wd} / (\varkappa + \lambda_{wd}),$$

il che stabilisce che la configurazione anisotropica amplifica il bias rispetto alla isotropica per quella specifica scelta di β^*. *Il Lemma 1 di LeJEPA formalizza questa osservazione affermando che, per $\lambda_{wd} > 0$, esiste sempre un compito downstream (una scelta di β^*)* per il quale la covarianza anisotropica produce un estimatore più distorto rispetto alla isotropica a parità di energia totale. L'intuizione geometrica è immediata: quando la distribuzione degli embedding è schiacciata lungo alcune direzioni, i parametri del probe che si allineano con quelle direzioni vengono sovrastimati dalla regolarizzazione, che non distingue tra dimensioni con varianza diversa.

Per la varianza, con $\lambda_{wd} = 0$ l'estimatore ai minimi quadrati ordinari è non distorto, e la sua covarianza è

$$\mathrm{Var}(\beta) = \sigma^2 (Z^T Z)^{-1} = \sigma^2 Q \Lambda^{-1} Q^T,$$

dove σ^2 è la varianza del rumore. La varianza totale diventa

$$\mathrm{tr}(\mathrm{Var}(\beta)) = \sigma^2 \mathrm{tr}(\Lambda^{-1}) = \sigma^2 \Sigma_{k=1}^K 1/\lambda_k.$$

Poiché la funzione $x \mapsto 1/x$ è strettamente convessa su $(0, \infty)$, la disuguaglianza di Jensen fornisce immediatamente

$$(1/K) \Sigma_k 1/\lambda_k \geq 1 / ((1/K) \Sigma_k \lambda_k) = K/\varkappa,$$

con uguaglianza se e solo se tutti gli autovalori sono uguali, cioè $\lambda_k = \varkappa$ per ogni k. Ne segue che

$\mathrm{tr}(\mathrm{Var}(\hat{\beta}))_{\mathrm{aniso}} \geq \mathrm{tr}(\mathrm{Var}(\hat{\beta}))_{\mathrm{iso}} = \sigma^2 K^2 / (N \varkappa)$,

e la disuguaglianza è stretta ogni volta che la distribuzione è anisotropica. Il Lemma 2 di LeJEPA: l'embedding isotropico minimizza la varianza totale dell'estimatore lineare a parità di energia totale della distribuzione.

Il passo successivo è l'identificazione della forma funzionale esatta della distribuzione ottimale, non solo della sua struttura di covarianza. Per il probing non-lineare tramite k-NN radiale, il bias del predittore in un punto di query q ha la forma

$\mathrm{Bias}(q) = r_0^2 / (d + 2) \cdot [\nabla \eta(q)^{\top} \nabla \log p_Z(q) + (1/2) \Delta \eta(q)] + o(r_0^2)$,

dove r_0 è il raggio della sfera di vicinato, $d = K$ la dimensione dell'embedding, $\eta : \mathbb{R}^K \to \mathbb{R}$ la funzione target downstream, e $\Delta\eta$ il laplaciano di η. Integrando il bias quadratico atteso rispetto alla distribuzione p_Z e su una famiglia di funzioni target η con gradiente isotropico $\mathbb{E}[\nabla\eta(z)\ \nabla\eta(z)^{\top}] = \tau^2_g I_K$, si ottiene

$\mathbb{E}_Z[\mathrm{Bias}(Z)^2] = (r_0^2 / (K + 2))^2 \cdot \tau^2_g \cdot J(p_Z) + O(r_0^4)$,

dove $J(p_Z) = \int \|\nabla \log p_Z(z)\|^2 p_Z(z)\, dz$ è il funzionale di Fisher information. Si tratta ora di minimizzare $J(p_Z)$ soggetto al vincolo scalare sulla covarianza. Poiché il bound di Cramér-Rao stabilisce che, per qualsiasi densità p_Z a covarianza Σ,

$J(p_Z) \geq \mathrm{tr}(\Sigma^{-1})$,

con uguaglianza se e solo se $p_Z = \mathcal{N}(0, \Sigma)$ (la gaussiana è l'unica distribuzione a saturare il bound di Cramér-Rao), il problema si riduce a minimizzare $\mathrm{tr}(\Sigma^{-1})$ soggetto al vincolo scalare. Con il vincolo di traccia $\mathrm{tr}(\Sigma) = \varkappa$, la disuguaglianza di Cauchy-Schwarz dà

$\sum_k 1/\lambda_k \cdot \sum_k \lambda_k \geq K^2$,

da cui $\mathrm{tr}(\Sigma^{-1}) \geq K^2/\varkappa$, con uguaglianza se e solo se tutti gli autovalori sono uguali a $\varkappa/K$, cioè $\Sigma = (\varkappa/K) I_K$. Per vincoli di tipo Frobenius $\|\Sigma\|_F = c$, le condizioni di Lagrange conducono agli stessi $\lambda_k = c/\sqrt{K}$ per ogni k. Per il vincolo sul raggio spettrale $\max_k \lambda_k \leq r$, il minimo è raggiunto con $\Sigma = r I_K$. In tutti i casi, la soluzione ottimale è la gaussiana isotropica $\mathcal{N}(0, s I_K)$ con s scelto per soddisfare il vincolo.

Questo risultato è enunciato nel Teorema 1 di LeJEPA: la distribuzione gaussiana isotropica è l'unico minimizzatore del bias integrato quadratico del probing k-NN (e del probing kernel) tra tutte le densità di probabilità con vincolo scalare sulla covarianza. La medesima conclusione vale per la regressione kernel di Nadaraya-Watson, dove il bias in un punto di query q ha la forma

$\mathrm{Bias}(q) = (h^2 \mu_2(K)) / 2 \cdot [\Delta\eta(q) + 2 \nabla\eta(q)^{\top} \nabla \log p_Z(q)] + o(h^2)$,

con h il bandwidth del kernel e $\mu_2(K)$ il secondo momento del kernel K. L'integrazione del bias quadratico soggetto alla funzione target nella classe $\mathcal{M}(L, B) = \{m \in C^2(\mathbb{R}^d) : \|\nabla m(z)\| \leq L, |\Delta m(z)| \leq B\}$ conduce allo stesso funzionale $J(p_Z)$, e la medesima argomentazione conclude che la gaussiana isotropica minimizza il worst-case bias integrato.

La portata di questi risultati va ben oltre la tecnica matematica. Per la prima volta nella storia dell'apprendimento autosupervisionato, si dispone di un principio normativo — non prescrittivo a priori né post-hoc giustificativo di soluzioni già trovate — che dice esattamente quale struttura geometrica lo spazio delle rappresentazioni deve avere per essere ottimale. Le euristiche precedenti (stop-gradient, teacher-student, whitening esplicito, coppie negative) possono essere tutte reinterpretate come approssimazioni imperfette del meccanismo che garantisce tale ottimalità: esse funzionano nella misura in cui inducono empiricamente distribuzioni di embedding che si avvicinano alla gaussiana isotropica, falliscono nella misura in cui se ne discostano.

III.5.3 — SIGReg: la regolarizzazione gaussiana isotropica per via di proiezioni stocastiche

Stabilito che la distribuzione ottimale è la gaussiana isotropica, resta il problema di realizzarla in modo scalabile e differenziabile durante l'addestramento. Le soluzioni ovvie — confrontare direttamente la distribuzione empirica degli embedding con una gaussiana isotropica tramite KL divergenza o MMD — sono o non differenziabili o computazionalmente proibitive (complessità quadratica nel numero di campioni). LeJEPA introduce a questo scopo la Sketched Isotropic Gaussian Regularization (SIGReg), un obiettivo che combina garanzie statistiche rigorose con complessità lineare nel numero di campioni e nel numero di dimensioni.

Il punto di partenza teorico è il teorema di Cramér-Wold nella versione iperferica: due vettori aleatori $X, Y \in \mathbb{R}^d$ hanno la medesima distribuzione se e solo se $\langle u, X\rangle$ e $\langle u, Y\rangle$ hanno la medesima distribuzione per ogni $u \in \mathcal{S}^{d-1}$. Questo teorema, dimostrato nel Lemma 3 di LeJEPA, riduce il problema del confronto multivariato a un insieme (tipicamente non numerabile) di confronti univariati. Il Teorema 2 di LeJEPA raffina ulteriormente questo risultato: data una famiglia finita di direzioni $A = \{a_1, \ldots, a_M\} \subset \mathcal{S}^{K-1}$, il test statistico

$$T_A(\{f_\theta(x)\}_{n=1}^N) = \max_{a \in A} T(\{a^T f_\theta(x)\}_\{n=1\}^N)$$

è un test valido per l'ipotesi nulla multivariata $P_\theta = Q$ nell'accezione statistica precisa: il livello α è controllato asintoticamente e la potenza converge a 1 quando $P_\theta \neq Q$, a condizione che le direzioni abbiano mesh decrescente a zero (o che vengono ricampionate durante l'addestramento). Ciò significa che, nella pratica dell'addestramento neurale, usare alcune centinaia di direzioni ricampionate a ogni mini-batch equivale, grazie all'accumulo durante il training, a coprire la sfera in

modo sempre più denso — il Teorema 5 di LeJEPA quantifica questa efficacia attraverso i coefficienti di regolarità di Sobolev dell'embedding.

SIGReg è definita formalmente come media (anziché massimo, per evitare gradienti sparsi) dei test statistici direzionali:

SIGReg_T(A, {f_θ(x)}{n=1}^N) = (1/|A|) Σ{a ∈ A} T({a^T f_θ(x)}_{n=1}^N),

dove T è un test statistico di bontà di adattamento a una gaussiana standard applicato alla variabile proiettata scalare. La scelta del test T è non banale: i test basati sui momenti (Jarque-Bera e la sua estensione) soffrono di instabilità dei gradienti perché il momento di ordine k ha gradiente con norma proporzionale a O(k) e varianza del gradiente proporzionale a O(k² m_{2(k−1)}), il che rende il training instabile per k ≥ 3 su distribuzioni con code pesanti. I test basati sulla CDF (Cramér-von Mises, Anderson-Darling, Watson) richiedono l'ordinamento dei campioni, operazione non parallela per natura che introduce sincronizzazioni costose su cluster multi-GPU e che genera gradienti non differenziabili senza apposite tecniche di rilassamento.

La scelta raccomandata in LeJEPA è il test di Epps-Pulley, basato sulla funzione caratteristica empirica:

EP = N ∫_{-∞}^{∞} |φ̂_X(t) − φ(t)|² w(t) dt,

dove φ̂_X(t) = (1/n) Σ_{j=1}^n e^{itX } è la funzione caratteristica empirica, φ(t) = e^{−t²/2} è la funzione caratteristica della gaussiana standard, e w(t) = e^{−t²/σ²} è una finestra gaussiana con σ = 1 di default. Questo test è naturalmente differenziabile (la funzione caratteristica empirica è una media di esponenziali complessi), distribuibile su GPU multi-nodo tramite l'operazione all_reduce (che calcola medie su processi), e ha gradienti uniformemente limitati come dimostra il Teorema 4 di LeJEPA:

$|\partial \mathrm{EP}(a)/\partial z_i| \le 4\sigma^2/N$, $|\partial^2 \mathrm{EP}(a)/\partial z_i^2| \le C\sqrt{\pi}\,\sigma^3 / (2N)$,

per qualche costante C. Questa limitazione uniforme dei gradienti — indipendente dai valori assunti dagli embedding — garantisce stabilità dell'addestramento per qualsiasi distribuzione di input, in netto contrasto con i test basati sui momenti che possono produrre gradienti esplosivi quando gli embedding hanno code pesanti.

In pratica, l'integrale del test di Epps-Pulley viene approssimato con la regola trapezoidale su una griglia di punti {t_1, …, t_J} con J = 17 e dominio [−5, 5]:

EP ≈ Σ_{j=1}^J ω_j |φ̂_X(t) − φ(t)|² w(t),

dove ω_j sono i pesi della quadratura trapezoidale. L'errore di quadratura è dimostrato essere O(1/N) per la gaussiana standard e converge rapidamente anche per distribuzioni bimodali (come 0.5·$\mathcal{N}$(−2, 0.25) + 0.5·$\mathcal{N}$(2, 0.25)) e distribuzioni a coda pesante (Student-t con ν = 3). La complessità computazionale dell'intero

SIGReg — incluse le M proiezioni, il calcolo delle funzioni caratteristiche empiriche e la valutazione della quadratura — è O(N · M · J), lineare in tutti e tre i fattori; con N = 8192, M = 1024, J = 17, il tempo di esecuzione misurato su una GPU Tesla V100 è circa 6 ms per forward e backward, del tutto trascurabile rispetto al costo dell'encoder e del predittore.

La questione della maledizione della dimensionalità merita una discussione dedicata. Uno dei principali ostacoli alla verifica statistica in alta dimensione è che il numero di direzioni necessario per coprire adeguatamente $\mathcal{S}^{K-1}$ cresce esponenzialmente con K. Il Teorema 5 di LeJEPA dimostra che, se la densità p_θ ha regolarità di Sobolev α sulla sfera (ovvero appartiene a $H^\alpha(\mathcal{S}^{K-1})$), allora l'errore atteso del test direzionale su M direzioni è limitato da

$$\mathbb{E}_a\left[\int_{\mathbb{R}} |\varphi_a(t) - \varphi_{\mathcal{N}}(t)|^2 \, dt\right] \leq C(K, a) \cdot |A|^{-2a/(K-1)} \cdot \|p_\theta\|^2_{H^\alpha(\mathcal{S}^{K-1})},$$

dove $C(K, \alpha) = 2^{2\alpha} \pi^{(K-1)/2} \Gamma(\alpha + (K-1)/2) / [(K-1) \Gamma(\alpha) \Gamma((K-1)/2)]$ è una costante esplicita che dipende da K e α. Il bound decade a zero come $|A|^{-2\alpha/(K-1)}$, il che significa che per α grande (distribuzione liscia) il numero di direzioni necessario per una data accuratezza ε scala come $O(K/\varepsilon^{1/(2\alpha/(K-1))}) = O(K \cdot \varepsilon^{-(K-1)/(2\alpha)})$. Per reti neurali profonde, la distribuzione degli embedding tende a diventare liscia già nelle prime fasi dell'addestramento (grazie alla regolarizzazione implicita della discesa del gradiente), il che corrisponde ad α grande e rende il fattore esponenziale nella dimensione molto più favorevole che nel caso peggiore. Inoltre, e questo è l'argomento decisivo dal punto di vista pratico, il ricampionamento delle direzioni a ogni mini-batch moltiplicato il numero di iterazioni di training produce una copertura cumulativa della sfera che cresce linearmente nel tempo, compensando la limitatezza di M per ogni singolo batch.

III.5.4 — LeJEPA: l'architettura e la sua perdita composita

Il framework LeJEPA si ottiene combinando la perdita predittiva standard di JEPA con SIGReg applicata agli embedding di ciascuna vista:

$$\mathcal{L}_{\text{LeJEPA}}(\{x_{,v}\}_{B,V}) = \lambda/V \cdot \Sigma_{v=1}^{V} SIGReg(\{z_{n,v}\}_{n=1}^{B}) + (1-\lambda)/B \cdot \Sigma_{=1}^{B} \mathcal{L}_{pred}^{(Vg)}(\{z_{n,v}\}_{v=1}^{V}),$$

dove $\{z_{n,v}\}$ sono gli embedding per l'n-esimo campione e la v-esima vista, $\lambda \in (0,1)$ è l'unico iperparametro di bilanciamento, e $\mathcal{L}_{pred}$ è la perdita predittiva. Quest'ultima viene sviluppata esplicitamente come la distanza quadratica media tra gli embedding di ciascuna vista e la media degli embedding delle viste globali:

$$\mathcal{L}_{pred}^{(Vg)}(\{z_{n,v}\}_{v=1}^{V}) = (1/V) \Sigma_{v=1}^{V} \|\mu - z_{n,v}\|^2_2,$$

dove $\mu = (1/Vg) \Sigma_{v=1}^{Vg} z_{n,v}$ è la media degli embedding delle Vg viste globali. La derivazione algebrica della forma equivalente — che la media

dell'ℒ_pred sulle viste globali eguaglia la norma della differenza tra μ e ogni z_{n,v} — si dimostra osservando che

(1/Vg) $\Sigma_{v=1}$^{Vg} (1/V) $\Sigma_{v'=1}$^V $\|z_{n,v} - z_{n,v'}\|^2_2 = \|\mu\|^2_2 - 2\mu$ ^T ((1/V) $\Sigma_{v'}$ z_{n,v'}) + (1/Vg) $\Sigma_v \|z_{n,v}\|^2_2$,

e poiché (1/Vg) $\Sigma_v \|z_{n,v}\|^2_2 = \|\mu\|^2_2$ (identità verificata espandendo $\|\mu\|^2_2 = \|(1/Vg) \Sigma_v z_{n,v}\|^2_2$ e usando l'indipendenza tra viste), si ottiene la forma compatta (1/V) $\Sigma_{v'} \|\mu - z_{n,v'}\|^2_2$.

L'intero sistema richiede meno di cinquanta righe di PyTorch per la parte non-standard (SIGReg e perdita predittiva), non impiega stop-gradient, non usa reti teacher-student, non richiede schedulatori degli iperparametri specifici per il dataset, ed è direttamente compatibile con il Distributed Data Parallel (DDP) di PyTorch grazie alla struttura di SIGReg come media di statistiche monodimensionali sincronizzabili tramite all_reduce. La stabilità dell'addestramento è documentata empiricamente fino a ViT-g/14 con 1,8 miliardi di parametri su ImageNet-1K, dove la curva di perdita rimane monotona decrescente senza collasso o divergenza durante tutte le 72 epoche di pretraining.

Un risultato particolarmente degno di nota riguarda la relazione tra la perdita di training di LeJEPA e le prestazioni downstream. Per i metodi di apprendimento autosupervisionato standard, questa relazione è sistematicamente bassa o addirittura inversa: perdite di training decrescenti non garantiscono rappresentazioni migliori, il che rende necessario il monitoraggio attraverso valutazioni supervisionate costose. LeJEPA rompe con questa prassi: la correlazione di Spearman tra la perdita di training e l'accuratezza di linear probing su ImageNet-1K con ViT-base/8 è 94.52%, e sale fino a 99% raffinando il coefficiente di allineamento α nella legge di scala

C^(α) = ϱ_s(perdita_training / λ^α, test_accuracy),

con $\alpha \approx 0.4$ in modo robusto su tutti i dataset e architetture testati. L'interpretazione teorica di questo risultato è la seguente: la perdita di LeJEPA è una somma di due termini la cui traiettoria congiunta nello spazio bidimensionale (SIGReg, perdita predittiva) descrive frontiere di pareto che sono linearmente allineate con le prestazioni downstream, così che il valore combinato — pesato da λ — è informativo della qualità delle rappresentazioni. Questo trasforma LeJEPA in uno strumento di selezione del modello senza etichette (label-free model selection), una capacità che non aveva precedenti nella letteratura SSL per immagini.

Le prestazioni empiriche confermano le previsioni teoriche in modo convincente. Con pretraining in-domain su Galaxy10 (11.000 immagini di morfologia galattica) e linear probing con backbone congelato, architetture di dimensione contenuta come ResNet-34 e ConvNeXt-V2 Nano addestrate con LeJEPA superano DINOv2 e DINOv3 (addestrati su 142M e 1,7B immagini di immagini naturali rispettivamente) in tutti i regimi di dati da 1-shot a supervisione completa. Questo

dimostra che l'apprendimento autosupervisionato in-domain su dati specializzati, quando fondato su principi rigorosi come LeJEPA, può superare il transfer learning da modelli frontier di scala enormemente maggiore — un risultato di considerevole importanza pratica per domini scientifici e industriali dove i dati etichettati scarseggiano e i dati non etichettati sono disponibili in quantità limitata ma specializzata.

Su benchmark di transfer learning few-shot rispetto a I-JEPA (ViT-H, 632M parametri, 300 epoche su IN-1K), LeJEPA con ViT-L (304M parametri, soli 100 epoche) raggiunge prestazioni comparabili o superiori su 7 degli 8 dataset testati (DTD, aircraft, cars, CIFAR-10, CIFAR-100, flowers102, food101, pets), con vantaggi particolarmente marcati sui compiti di classificazione fine-grained (flowers102: 92.53% vs 88.24% per I-JEPA a 10-shot). Il vantaggio in epoche di addestramento — un fattore tre rispetto a I-JEPA — si traduce direttamente in un vantaggio computazionale che rende praticabile il pretraining in domini specializzati su hardware non di datacenter.

L'analisi delle rappresentazioni attraverso la PCA delle feature dell'ultimo layer di LeJEPA su ViT-Large addestrato su ImageNet-1K rivela la comparsa di strutture semantiche emergenti non supervisionate: i primi tre componenti principali codificano sistematicamente la separazione primo piano-sfondo, il raggruppamento percettivo degli oggetti e la struttura di illuminazione della scena. Questo fenomeno emergente — già documentato per DINO ma qui ottenuto con un obiettivo più semplice e teoricamente fondato — conferma che la gaussiana isotropica come distribuzione target produce rappresentazioni con proprietà strutturali profonde, non soltanto metriche di classificazione superiori.

III.5.5 — Dalle euristiche come casi speciali: il recupero di VICReg e la gerarchia metodologica

Il percorso teorico di LeJEPA non si conclude con la derivazione della distribuzione ottimale e la costruzione di SIGReg come meccanismo per raggiungerla. Un terzo risultato, di importanza non inferiore ai precedenti, riguarda il posizionamento di LeJEPA all'interno del paesaggio metodologico dell'apprendimento autosupervisionato: la dimostrazione formale che i metodi precedenti — VICReg, SimCLR, BYOL, DINO — sono, da un punto di vista teorico, casi speciali o approssimazioni imperfette dello stesso principio ottimale, e che le loro euristiche anti-collasso sono tentativi parziali di imporre la stessa struttura geometrica che LeJEPA impone in modo rigoroso e completo. Questo terzo risultato non è una critica retorica ai metodi precedenti: è una riorganizzazione della mappa teorica che attribuisce a ciascun metodo il suo posto nella gerarchia delle approssimazioni al vero ottimo.

Il punto di partenza è il Teorema 3 di LeJEPA, che stabilisce un risultato di carattere negativo fondamentale: il matching di un numero finito K di momenti non è sufficiente a garantire l'uguaglianza distributiva. Formalmente, sia $T_K(\{x_n\}) = \Sigma_{=1}^{k} c_k (m_k(P_\theta^{(a)}) - m_k(Q^{(a)}))^2$ una statistica che misura la

discrepanza pesata tra i primi K momenti della distribuzione proiettata empirica P_θ^(a) e i corrispondenti momenti della gaussiana standard Q^(a), con c_k > 0 per ogni k. La tesi è che la minimizzazione di T_K non implica P_θ^(a) = Q^(a). La dimostrazione è costruttiva e istruttiva: dati K+2 punti distinti x_0, ..., x_{K+1} ∈ ℝ, la mappa A : ℝ^{K+2} → ℝ^{K+1} definita da (Ap)_r = Σ_j p_j x_j^r per r = 0, ..., K ha rango al più K+1 per il teorema di Vandermonde applicato all'algebra lineare, dunque il suo nucleo ker(A) è non banale. Scegliendo v ∈ ker(A) \ {0} con componenti di segno alterno in modo che p^+ := p + εv e p^- := p - εv siano entrambe distribuzioni di probabilità strettamente positive, si ottiene A p^+ = A p^-, ovvero due distribuzioni distinte con identici primi K+1 momenti. Ne segue che esiste sempre una famiglia continua di distribuzioni non gaussiane che minimizzano qualsiasi obiettivo basato su K momenti finiti: queste distribuzioni costituiscono le soluzioni di shortcut inaccessibili ai metodi basati sul matching di momenti.

La conseguenza immediata del Teorema 3 per la progettazione di SIGReg è che il test T deve utilizzare tutte le informazioni distribuzionali, non solo un numero finito di statistiche sufficienti. La funzione caratteristica empirica $\hat{\phi}_X(t) = (1/n) \Sigma_j e^{itX_j}$ risponde esattamente a questo requisito: il test di Epps-Pulley misura la distanza ponderata tra la funzione caratteristica empirica e quella della gaussiana standard sull'intero asse reale, catturando quindi l'informazione di tutti i momenti simultaneamente e senza l'instabilità che affligge le stime dei momenti di ordine elevato. Il Teorema di Inversione di Lévy garantisce che l'uguaglianza delle funzioni caratteristiche è condizione necessaria e sufficiente per l'uguaglianza distributiva, il che rende il test di Epps-Pulley statisticamente completo nel senso tecnico preciso: la sua minimizzazione esclude tutte le soluzioni di shortcut che i test basati su momenti finiti lascerebbero aperte.

Il Teorema di Recupero di VICReg, dimostrato nella Sezione 5.2 di LeJEPA, istanzia esplicitamente il Teorema 3 nel caso di VICReg. Si consideri la statistica $T_{VIC}(\{x_n\}) = \text{mean}(\{x_n\})^2 + (\text{std}(\{x_n\}) - 1)^2$, ovvero la somma della media quadratica e della deviazione quadratica dalla varianza unitaria della variabile scalare proiettata: questa è una statistica di test a due momenti (primo e secondo). Sostituendo T_VIC come test T nella definizione di SIGReg e prendendo il limite per $|A| \to \infty$ con direzioni uniformemente distribuite su S^{K-1}, si verifica che $SIGReg_{T_{VIC}}$ impone in attesa $E[Z] = 0$ e $\text{Cov}(Z) = I_K$ — esattamente i vincoli di VICReg. La dimostrazione si articola in due passi. Per la media: $E[\langle a, Z\rangle] = 0$ per ogni direzione unitaria a implica, per linearità dell'attesa, $a^T E[Z] = 0$ per ogni $a \in S^{K-1}$, dunque $E[Z] = 0$ scegliendo $a = e_i / \|e_i\|$ per ogni coordinata. Per la covarianza: $\text{Var}(\langle a, Z\rangle) = 1$ per ogni a implica $a^T \text{Cov}(Z)\, a = 1$ per ogni vettore unitario, da cui — valutando sulle direzioni canoniche e_i (che danno $\Sigma_{ii} = 1$) e sulle direzioni $(e_i + e_j)/\sqrt{2}$ (che danno $1/2(\Sigma_{ii} + 2\Sigma_{ij} + \Sigma_{jj}) = 1$, ovvero $\Sigma_{ij} = 0$ per $i \neq j$) — si ricava $\text{Cov}(Z) = I_K$. VICReg è dunque un caso degenere di LeJEPA: l'obiettivo di regolarizzazione di VICReg corrisponde esattamente a SIGReg con T = T_VIC, ovvero con un test a

due momenti che, per il Teorema 3, non esclude le soluzioni di shortcut distribuzionali.

La stessa logica si applica, con modulazioni diverse, alle altre principali famiglie di metodi SSL. In BYOL e SimSiam il stop-gradient sul branch target impedisce il collasso completo ma non garantisce la forma distribuzionale degli embedding: la rete può convergere a distribuzioni con covarianza anisotropica o code pesanti, fenomeno documentato empiricamente come collasso dimensionale da Jing et al. (2021). In DINO e DINOv2 la rete teacher con aggiornamento EMA svolge una funzione analoga di stabilizzazione; LeJEPA fornisce retrospettivamente una giustificazione teorica parziale di questo meccanismo: l'aggiornamento EMA è un'approssimazione dello Stochastic Weight Averaging che migliora la stima della distribuzione target degli embedding, ma non è un meccanismo che garantisce per costruzione l'isotropia. La whitening esplicita in W-MSE e SwAV impone l'isotropia della covarianza, avvicinandosi maggiormente alla struttura ottimale; tuttavia, whitening sulla sola covarianza equivale a imporre solo i secondi momenti degli embedding, lasciando non vincolata la forma distribuzionale completa. In tutti questi casi, il meccanismo empirico funziona nella misura in cui spinge la distribuzione degli embedding verso configurazioni più prossime alla gaussiana isotropica, e presenta instabilità o dipendenza da iperparametri nella misura in cui tale pressione è incompleta.

La differenza qualitativa tra LeJEPA e tutti i metodi precedenti può essere formulata con precisione in termini di eliminazione del collasso. Per qualsiasi configurazione di embedding in collasso — sia collasso completo (tutti gli embedding identici, distribuzione degenere in un punto) sia collasso dimensionale (embedding confinati in un sottospazio di dimensione $d' < K$) — SIGReg con Epps-Pulley assegna una perdita strettamente positiva con gradiente non nullo che spinge il sistema fuori dalla configurazione collassata. La ragione è diretta: qualsiasi distribuzione collassata si discosta dalla gaussiana isotropica in modo rilevabile da almeno una proiezione su S^{K-1}. Una distribuzione concentrata in un punto ha varianza zero su ogni proiezione (la gaussiana standard ha varianza unitaria), producendo un valore del test di Epps-Pulley proporzionale a N. Una distribuzione su un sottospazio d'-dimensionale ha varianza zero su tutte le $K - d'$ direzioni ortogonali al sottospazio; l'integrale della funzione caratteristica lungo tali direzioni si discosta sistematicamente da quello della gaussiana, generando un segnale di gradiente che gonfia la varianza nelle direzioni collassate. La limitazione uniforme dei gradienti di Epps-Pulley ($|dEP/dz_i| \leq 4\sigma^2/N$, indipendentemente dalla distribuzione corrente degli embedding) garantisce che questa spinta sia numericamente stabile anche in regime di collasso profondo, evitando sia i gradienti esplosivi dei test basati sui momenti sia i gradienti sparsi dei test basati sulla CDF.

La portata epistemica di questa riorganizzazione metodologica va oltre il tecnico. Ciò che LeJEPA realizza è una trasformazione del tipo di giustificazione disponibile per i metodi SSL: dalla giustificazione post-hoc — 'questo funziona empiricamente, e qui si propone una spiegazione teorica ex post' — alla

giustificazione a priori — 'questo metodo segue necessariamente da principi normativi dimostrabili, e qualsiasi metodo che funzioni empiricamente deve implicarlo come caso speciale o approssimazione'. Il Teorema 3 spiega perché VICReg è instabile su certi dataset; il recupero formale di VICReg come caso degenere di SIGReg spiega perché VICReg funziona bene dove funziona; il Teorema 1 spiega perché qualsiasi metodo che imponga la gaussiana isotropica — anche per vie euristiche — si avvicina all'ottimo downstream. Questa struttura gerarchica di giustificazioni teoriche, in cui i risultati positivi (cosa funziona e perché) e i risultati negativi (cosa non funziona e perché) sono derivati dallo stesso principio normativo, costituisce il tipo di comprensione che la sezione I.3 di questa monografia ha identificato come condizione necessaria per il progresso scientifico non accidentale. Il passaggio da euristiche a teoremi non è un abbellimento formale: è il segnale che il campo ha raggiunto la maturità teorica sufficiente per progredire per via deduttiva anzié per via meramente sperimentale.

III.6 — V-JEPA: estensione ai video

V-JEPA (Video-JEPA), presentato da Meta AI nel 2024, estende I-JEPA al dominio video applicando il principio della predizione latente a sequenze di frame temporali. L'estensione non è banale perché introduce una dimensione aggiuntiva di struttura — quella temporale — che può essere sfruttata attraverso strategie di mascheramento specifiche.

Formalmente, sia $\mathbf{X} = (X_1, X_2, ..., X_T) \in \mathbb{R}^{(T\times H\times W\times 3)}$ un clip video di T frame. V-JEPA tratta i frame come una sequenza 3D di patch spaziotemporali: ciascuna patch è un cuboid di dimensioni $P\times P\times \tau$ pixel ($P=16$, $\tau=2$ tipicamente), e l'intera sequenza viene tokenizzata in $N_st = (T/\tau) \times (H/P) \times (W/P)$ token. L'architettura è analoga a I-JEPA con un context encoder, un target encoder EMA e un predictor, ma il mascheramento avviene nello spazio spaziotemporale.

La strategia di mascheramento di V-JEPA è particolarmente interessante. Si usano due varianti principali:

Il **mascheramento spaziotemporale completo** seleziona un sottoinsieme casuale di frame interi come target e usa il resto come contesto. Questo compito corrisponde a predire frame futuri (o passati) dato un contesto parziale, il che richiede la comprensione della dinamica della scena.

Il **mascheramento spaziale per frame** seleziona la stessa regione spaziale in tutti i frame come target, lasciando il contesto spaziale di ogni frame intatto. Questo compito è strutturalmente diverso: richiede al modello di integrare informazioni temporali per "capire" cos'è presente nella regione mascherata attraverso la sua storia temporale.

La perdita di V-JEPA è analoga a quella di I-JEPA:

$$\mathcal{L}_VJEPA(\theta, \varphi) = \mathbb{E} [\Sigma_{(t,i) \in T} \| P_\varphi(E_\theta(X_C), (t, pos_i)) - sg(\bar{E}_\theta(X)_{t,i}) \|^2]$$

dove l'indice di target include sia il frame t che la posizione spaziale i.

L'analisi delle proprietà di V-JEPA rivela che le rappresentazioni apprese catturano informazioni sulla dinamica degli oggetti in modo significativamente migliore rispetto a rappresentazioni apprese su singoli frame: in probe tasks che richiedono la comprensione di motion, deformazione e interazione tra oggetti, V-JEPA supera sia I-JEPA fine-tuned su video che metodi contrastivi video come VideoMAE e TimeSformer. Questo suggerisce che la predizione latente spaziotemporale è una forma di supervisione intrinsecamente più ricca per il reasoning dinamico rispetto alla predizione a livello di frame isolato.

III.6.1 — Dalla percezione all'agentività: il salto architetturale di V-JEPA 2

V-JEPA 2, presentato da Meta AI nel corso del 2025 (Assran et al., 2025), non è una semplice revisione incrementale di V-JEPA ma una reinterpretazione strutturale del framework che introduce per la prima volta una componente di pianificazione agentica integrata con il modulo percettivo. Mentre V-JEPA era concepito esclusivamente come strumento di apprendimento di rappresentazioni — il suo fine era la qualità delle feature estratte, valutata attraverso probing supervisionato — V-JEPA 2 si propone esplicitamente come sistema capace di «comprendere e predire come il mondo fisico si evolve», ambizione che trascina con sé un apparato retorico di comprensione che il presente capitolo analizza criticamente nelle sue implicazioni epistemologiche (cfr. Capitolo XI) e che occorre qui radicare in una descrizione tecnica rigorosa.

L'architettura di V-JEPA 2 eredita il nucleo di V-JEPA — l'encoder contestuale, il target encoder EMA, il predittore spaziotemporale — ma lo aumenta con due componenti aggiuntive: un **modulo di condizionamento sull'azione** che consente al sistema di ricevere come input sequenze di azioni e di predire la rappresentazione latente dello stato futuro del mondo condizionatamente ad esse, e un **modulo di pianificazione** che utilizza il world model interno per simulare le conseguenze di sequenze di azioni alternative e selezionare quelle che minimizzano una funzione di costo definita esternamente. Questo schema corrisponde precisamente alla struttura del Model Predictive Control (MPC) analizzata nella sezione V.3, applicata non a uno spazio di stati fisici espliciti ma allo spazio latente dell'encoder.

III.6.2 — Formalismo dell'estensione agentica

Sia $\mathbf{X} = (X_1, \ldots, X_T) \in \mathbb{R}^{T \times H \times W \times 3}$ un clip video di T frame e sia $\mathbf{a} = (a_1, \ldots, a_{T-1}) \in \mathcal{A}^{T-1}$ la sequenza di azioni intraprese tra fotogrammi consecutivi, dove $\mathcal{A}$ è lo spazio delle azioni (vettori continui di dimensione d_a per manipolazione robotica, segnali discreti per ambienti di gioco). V-JEPA 2 apprende le funzioni seguenti.

L'encoder contestuale $\mathbf{E}_\theta : \mathbb{R}^{T \times H \times W \times 3} \to \mathbb{R}^{N_{st} \times d}$ è, come in V-JEPA, un Vision Transformer 3D che trasforma il clip in una sequenza di rappresentazioni spaziotemporali. La tokenizzazione procede attraverso cuboid

P×P×τ con P = 16 e τ = 2 nella configurazione standard, producendo N_st = (T/τ) × (H/P) × (W/P) token.

Il predittore condizionato sull'azione **P_φ** : ℝ^(N_C × d) × ℝ^(N_T × d_pos) × ℝ^(T_a × d_a) → ℝ^(N_T × d) estende il predittore di V-JEPA con un meccanismo di cross-attention sull'embedding dell'azione. Sia **a**_emb = W_a · **a** ∈ ℝ^(T_a × d) la proiezione lineare della sequenza di azioni nello spazio delle rappresentazioni, con W_a ∈ ℝ^(d × d_a) una matrice apprendibile. Il predittore modifica l'attenzione come segue:

H'_pred = MHSA(LN(H_context)) + CrossAttn(LN(H_context), **a**_emb)

dove CrossAttn applica attenzione multi-testa con le query derivate da H_context e le chiavi/valori derivati da **a**_emb. Questo meccanismo permette al predittore di modulare la propria risposta in funzione delle azioni intraprese, apprendendo implicitamente un modello della dinamica azione-stato nello spazio latente.

La perdita di addestramento di V-JEPA 2 comprende due termini. Il primo è il termine di predizione latente, identico nella forma a V-JEPA ma con il predittore condizionato:

$$\mathcal{L}_{pred}(\theta, \varphi) = \mathbb{E}_{X, a, M} \left[\Sigma_{(t,i) \in T} \| P_\varphi(E_\theta(X_C), (t, pos_i), \mathbf{a}) - sg(\bar{E}_{\bar{\theta}}(X)_{t,i}) \|^2 \right]$$

Il secondo è un termine opzionale di coerenza temporale che penalizza le variazioni eccessive della rappresentazione latente tra frame consecutivi non connessi da azioni di grande entità:

$$\mathcal{L}_{temp}(\theta) = \mathbb{E}_{X, a} \left[\Sigma_{t=1}^{T-1} \max(0, \| \bar{E}_{\bar{\theta}}(X)_t - \bar{E}_{\bar{\theta}}(X)_{t+1} \|^2 - \gamma \cdot \| a_t \|^2) \right]$$

dove γ è un iperparametro che scala l'ampiezza dell'azione. Questo termine incoraggia la smoothness della traiettoria latente in assenza di azioni di grande ampiezza, una proprietà utile per il planning su orizzonti lunghi.

La perdita complessiva è:

$$\mathcal{L}_{VJEPA2}(\theta, \varphi) = \mathcal{L}_{pred}(\theta, \varphi) + \lambda_{temp} \cdot \mathcal{L}_{temp}(\theta)$$

con λ_temp determinato attraverso validazione su dataset di robot manipulation.

III.6.3 — La pipeline di curazione dei dati e la sua struttura categoriale

L'aspetto più rilevante di V-JEPA 2 per la critica epistemologica sviluppata nel Capitolo XI — e il più sistematicamente trascurato nei commenti tecnici che si concentrano sulle innovazioni architetturali — è la pipeline di curazione del corpus di addestramento, descritta nella sezione 2.3 di Assran et al. (2025).

Il corpus primario di addestramento è YT1B, un insieme di circa un miliardo di clip video estratti dalla piattaforma YouTube. Questo corpus è «non curato e non

bilanciato», con le conseguenze documentate nella letteratura sullo squilibrio di distribuzione: predominanza di video di entertainment, turismo, cucina e sport nelle proporzioni che riflettono i gusti dell'utente medio della piattaforma, scarsità di video di fisica elementare, manipolazione di oggetti inanimati, interazioni causali non mediate dall'intenzione umana. Per compensare questo squilibrio, Assran et al. costruiscono una pipeline di curazione che filtra YT1B attraverso un retrieval basato su embedding: per ogni clip candidato, si calcola la distanza nel spazio delle embedding CLIP rispetto a clip di riferimento estratti dai dataset di benchmark, e si selezionano soltanto i clip la cui distanza non supera una soglia ε:

$$\text{YT1B_curato} = \{\, c \in \text{YT1B} : \exists\ r \in \mathcal{D}\text{_target t.c. dist_CLIP}(c, r) \leq \varepsilon \,\}$$

dove $\mathcal{D}$_target = Kinetics ∪ Something-Something-v2 ∪ COIN ∪ EpicKitchens è la distribuzione bersaglio composta dai training set di questi quattro dataset. La soglia ε è calibrata in modo che la distribuzione del corpus curato sia statisticamente vicina alla distribuzione di $\mathcal{D}$_target misurata attraverso i medesimi embedding CLIP.

La struttura di questa pipeline ha tre implicazioni che il Capitolo XI analizza nel dettaglio.

La prima implicazione è che i dataset di benchmark Kinetics, Something-Something-v2, COIN e EpicKitchens sono stati costruiti da esseri umani attraverso processi di annotazione esplicita: annotatori umani hanno definito le categorie di azione (400 categorie per Kinetics-400, 174 per Something-Something-v2), hanno selezionato i video rappresentativi di ciascuna categoria, hanno verificato la coerenza delle annotazioni. La struttura categoriale di questi dataset è pertanto una proiezione della struttura categoriale delle azioni nell'ontologia di un annotatore umano adulto, con tutte le contingenze culturali, linguistiche e percettive che ciò implica.

La seconda implicazione è che la curazione per prossimità nell'embedding CLIP propaga implicitamente non soltanto la distribuzione delle categorie di $\mathcal{D}$_target ma anche i bias rappresentativi di CLIP stesso, che è stato addestrato su coppie immagine-testo estratte dal web con la loro distribuzione di linguaggi, culture e rappresentazioni demografiche.

La terza implicazione, la più diretta per la critica epistemologica, è che il sistema viene poi valutato su benchmark estratti dalla medesima $\mathcal{D}$_target utilizzata come distribuzione bersaglio della curazione. La circolarità è formalmente misurabile: sia S_train il corpus YT1B_curato e sia S_test il test set dei benchmark; poiché YT1B_curato è stato filtrato per prossimità a $\mathcal{D}$_target e S_test è un sottoinsieme di $\mathcal{D}$_target, la distanza media dist_CLIP(S_train, S_test) è, per costruzione, inferiore alla distanza che si osserverebbe tra corpus non curati. Il sistema viene addestrato su video selezionati per assomigliare ai benchmark e poi valutato su quei benchmark: la circolarità non è un difetto accidentale ma una proprietà strutturale della pipeline.

III.6.4 — Evaluation su robot manipulation e limite della generalizzazione

V-JEPA 2 introduce rispetto a V-JEPA un'estensione verso compiti di robot manipulation, valutata su benchmark come MetaWorld e Calvin. In questi benchmark, un robot con braccio articolato deve eseguire sequenze di azioni di manipolazione — afferrare un oggetto, spostarlo in una posizione bersaglio, aprire un cassetto — e il sistema viene valutato sulla percentuale di task completati con successo.

I risultati di V-JEPA 2 su questi benchmark sono competitivi rispetto alle baseline considerate nel paper. La struttura dell'evaluation è però rivelatrice: sia in MetaWorld che in Calvin, l'ambiente di simulazione è altamente strutturato — oggetti geometricamente semplici, illuminazione uniforme, dinamica fisicamente deterministica, spazio delle azioni a bassa dimensionalità — e la distribuzione di addestramento dei task corrisponde quasi esattamente alla distribuzione di test. Quando le condizioni si discostano dalla distribuzione di addestramento — oggetti con texture non viste, illuminazione variabile, perturbazioni della dinamica fisica — le prestazioni degradano in modo marcato, un comportamento che nel paper viene denominato «distribution shift» e documentato in una sezione di ablation.

Questo comportamento è esattamente quello previsto dai teoremi di impossibilità del Capitolo XI: il sistema non ha appreso un modello causale della dinamica fisica ma ha appreso una funzione di corrispondenza tra osservazioni e azioni che è accurata nella distribuzione di addestramento e che decade fuori da essa perché non possiede i criteri di rilevanza necessari per identificare le strutture fisiche invarianti oltre la distribuzione osservata.

III.6.5 — Connessione con la Proposizione V.3 dell'AMI Framework

Nel framework AMI (Capitolo V), V-JEPA 2 occupa la posizione del Perception Module integrato con il World Model: l'encoder fornisce la rappresentazione latente dello stato del mondo $\hat{\mathbf{s}}_t = E_\theta(X_t)$, e il predittore condizionato sull'azione approssima la transizione di stato $\hat{\mathbf{s}}_{t+1} \approx P_\varphi(\hat{\mathbf{s}}_t, a_t)$. Questa struttura è formalmente identica al Recurrent State Space Model (RSSM) analizzato nella sezione IV.3, con la differenza che il modello di transizione è appreso implicitamente attraverso la predizione latente in V-JEPA 2, mentre nei modelli DreamerV2/V3 è appreso attraverso la massimizzazione dell'ELBO su una distribuzione posteriore esplicita.

La differenza non è meramente tecnica: il RSSM modella esplicitamente l'incertezza sullo stato futuro attraverso la distribuzione $p(z_{t+1} \mid h_{t+1})$ e la latente stocastica z_t, mentre V-JEPA 2 produce predizioni deterministiche dello stato futuro (la rappresentazione latente prevista) senza quantificare l'incertezza associata. Per il planning, questa mancanza è significativa: un agente che pianifica su orizzonti lunghi senza modellare l'incertezza non può distinguere tra stati del mondo in cui il proprio modello è affidabile e stati in cui l'errore di predizione

accumulato ha reso le proiezioni inaffidabili — la stessa limitazione identificata nella sezione X.3 come «quarta barriera» al framework AMI.

III.7 — MC-JEPA e l'estensione multi-scala

Una limitazione strutturale di I-JEPA nella sua formulazione originale è che opera su una singola scala spaziale: tutte le patch hanno la stessa dimensione P×P e il predictor opera a questa scala fissa. Tuttavia, le immagini naturali hanno strutture rilevanti a scale molto diverse — dalla texture fine di una superficie all'organizzazione globale di una scena — e una buona rappresentazione dovrebbe catturare tutte queste scale simultaneamente.

MC-JEPA (Multi-Crop JEPA) affronta questo problema adottando la strategia multi-crop di DINO e MSN: per ciascuna immagine si campionano viste a risoluzioni diverse — tipicamente due viste "globali" a risoluzione piena (224×224) e quattro o sei viste "locali" a risoluzione ridotta (96×96) — e il compito predittivo viene definito tra la vista locale (contesto) e la vista globale (target).

Formalmente, siano X^G e X^L rispettivamente una vista globale e una vista locale dello stesso campione, ottenute attraverso ritaglio e ridimensionamento. Si chiede al predictor di produrre, a partire dalle rappresentazioni latenti della vista locale, le rappresentazioni delle patch corrispondenti nella vista globale:

$$\mathcal{L}_{MCJEPA}(\theta, \varphi) = \mathbb{E}\left[\left\| P_\varphi(E_\theta(X^L), F(X^L, X^G)) - sg(\bar{E}_{\bar{\theta}}(X^G)_{aligned}) \right\|^2 \right]$$

dove F(X^L, X^G) è una funzione che calcola l'embedding posizionale delle patch di X^L proiettato nello spazio di X^G — ovvero tiene traccia di dove si trova la vista locale all'interno della vista globale — e aligned indica le patch della vista globale che corrispondono spazialmente alla vista locale.

Questo compito di predizione cross-scala ha una proprietà interessante: richiede al modello di "completare" la rappresentazione di una scena a partire da un dettaglio locale, che è essenzialmente un tipo di "ragionamento dal particolare al generale" — dal dettaglio alla struttura d'insieme. Le rappresentazioni risultanti tendono a essere più robuste rispetto a variazioni di scala e campo visivo, il che si traduce in migliori prestazioni su benchmark che richiedono localizzazione e detection a scale multiple.

III.8 — Analisi comparativa delle prestazioni e interpretazione

Le prestazioni di I-JEPA sui benchmark standard forniscono conferma empirica delle previsioni teoriche sviluppate nei capitoli precedenti, e vale la pena analizzarle in dettaglio per verificare la coerenza tra teoria e pratica.

Su ImageNet-1K con linear probing — il benchmark più indicativo della qualità delle rappresentazioni apprese, poiché consiste nel addestrare solo un classificatore lineare sulle rappresentazioni congelate del backbone — I-JEPA con ViT-H (632M parametri) raggiunge circa 81.9% top-1 accuracy, comparabile con MAE (83.1%

con ViT-H) ma inferiore a DINO v2 (86.7% con ViT-g). La differenza rispetto a DINOv2 è significativa ma non sorprendente: DINOv2 usa training su dati curati di grande scala (142M immagini) mentre I-JEPA nelle valutazioni standard usa ImageNet-1K (1.28M immagini), e soprattutto DINOv2 è un metodo di distillazione che sfrutta un modello teacher pre-addestrato, ovvero usa una forma di supervisione aggiuntiva.

Il dato più interessante riguarda i compiti che richiedono comprensione strutturale piuttosto che classificazione: su Counting (stimare il numero di oggetti in una scena) e Depth Estimation (stimare la profondità), I-JEPA supera sia MAE che BEiT nonostante le prestazioni inferiori in classificazione lineare. Questo è esattamente il pattern predetto dalla teoria: la predizione latente a blocchi costringe il modello a catturare strutture spaziali e relazioni tra regioni, il che è meno direttamente rilevante per la classificazione ma cruciale per il ragionamento strutturale.

Su Downstream tasks che richiedono fine-tuning completo, il divario rispetto a MAE si riduce: con fine-tuning di tutti i parametri su ImageNet-1K, I-JEPA-ViT-H raggiunge 86.9% contro 87.8% di MAE-ViT-H. Questa convergenza suggerisce che i vantaggi di I-JEPA sulle rappresentazioni siano parzialmente "lavabili via" durante il fine-tuning, il che è prevedibile: il fine-tuning adatta le rappresentazioni al compito target, riducendo l'importanza della qualità intrinseca delle rappresentazioni pre-addestrate.

Un'analisi interessante riguarda il sample efficiency: con quante epoche di pretraining I-JEPA raggiunge prestazioni comparabili a MAE? I risultati mostrano che I-JEPA raggiunge prestazioni competitive con circa 600 epoch contro le 1600 epoch standard di MAE — ovvero richiede circa 2.6 volte meno compute per raggiungere qualità simile. Questo si spiega con il fatto che la predizione latente è un segnale di apprendimento più "denso" di quello della ricostruzione sensoriale: ogni patch target fornisce un segnale semanticamente ricco piuttosto che un segnale di pixel ad alta varianza.

Questi risultati empirici acquisiscono ancora più valore quando vengono messi in relazione con le previsioni teoriche sviluppate nel Capitolo II. La sezione II.7 aveva argomentato che l'obiettivo di predizione latente produce encoder con Lipschitz constant inferiore rispetto agli encoder addestrati su ricostruzione sensoriale, il che dovrebbe tradursi in migliore calibrazione e migliore generalizzazione fuori distribuzione. I risultati su Counting e Depth Estimation — compiti che richiedono la comprensione di relazioni strutturali tra regioni dell'immagine piuttosto che la classificazione di feature locali — sono coerenti con questa previsione: l'encoder JEPA cattura strutture geometriche e relazionali in modo più affidabile di MAE perché il suo obiettivo lo spinge verso quella struttura, non verso la fedeltà sensoriale.

Un aspetto che merita una riflessione aggiuntiva riguarda la natura del benchmark di linear probing su ImageNet. Sebbene questo benchmark sia diventato lo

standard de facto per la valutazione delle rappresentazioni autosupervisionato, esso misura specificamente la separabilità lineare delle rappresentazioni rispetto alle classi di ImageNet — un sottoinsieme molto ristretto delle proprietà che una rappresentazione dovrebbe avere per supportare ragionamento generale. Un sistema che eccelle in linear probing può avere rappresentazioni organizzate in modo utile per la classificazione ma inutile per la composizione, per il ragionamento controfattuale, o per la pianificazione su sequenze di azioni. La valutazione delle rappresentazioni di I-JEPA su compiti strutturali come il conteggio degli oggetti e la stima della profondità è un passo nella direzione giusta: misura proprietà geometriche e relazionali piuttosto che categoriali, ed è per questo motivo teoricamente più informativa sulla qualità delle rappresentazioni nel senso rilevante per AMI.

III.8.B — Benchmark comparativi di GMJE su CIFAR-10 e MNIST

I risultati empirici di Huang (2026) su CIFAR-10 e MNIST integrano l'analisi comparativa di I-JEPA sviluppata nella sezione III.8 con una prospettiva complementare: non la qualità delle rappresentazioni in assoluto ma l'efficienza del meccanismo di apprendimento contrastivo. L'esperimento chiave (sezione 5.2.1) confronta due sistemi con backbone ResNet-18 identico, stessa augmentation pipeline SimCLR, stesso EMA momentum su CIFAR-10 per 1000 epoche, con la sola differenza nel meccanismo di aggiornamento della memory bank (M=256 slot): MoCo v2 FIFO vs SMC-GMJE particle filter. MoCo v2 raggiunge 78.08% di linear probing, SMC-GMJE 81.99% (+3.91%). L'ESS (Effective Sample Size = $1/\Sigma(W^{(m)})^2$) del filtro SMC scende bruscamente intorno all'epoca 500, segnalando la concentrazione delle probabilità su un sottoinsieme piccolo di hard negatives che vengono clonati dal SIR. Questa concentrazione è esattamente il comportamento desiderato: la memory bank diventa un repertorio di campioni massimamente informativi per la discriminazione.

Il benchmark più largo (sezione 5.2.2, 200 epoche) colloca SMC-GMJE (74.52%) sotto SimCLR (81.50%), BYOL (79.04%) e MoCo v2 (79.02%) con la stessa backbone. Questa posizione è coerente con lo scenario previsto: SMC-GMJE è la prima implementazione principiata del framework GMJE e non ha beneficiato dei decenni di tuning empirico dei metodi baseline. Il pattern — metodo fondazionalmente corretto ma inizialmente sotto-performante rispetto a baseline mature — si è già verificato per I-JEPA rispetto a DINOv2 nelle prime valutazioni (81.9% vs 86.7% in linear probing). I risultati sull'esperimento di lunga durata con M=256 (sezione 5.2.1) mostrano che il vantaggio di SMC-GMJE cresce con il numero di epoche e con la scarsità della memory bank — esattamente il regime più rilevante per applicazioni con limiti di memoria.

Per i benchmark generativi su MNIST (sezione 5.3), GMJE parametrico (K=50 componenti) produce campioni latenti con prototype spread L2 = 1.43 rispetto a 0.55 di SimCLR+GMM e N/A (non applicabile) di GJE unimodale. La media $\log|\Sigma_k|$ (volume dei componenti) è −621.58 per GMJE vs −636.15 per GJE vs −878.99 per SimCLR+GMM: GMJE mantiene la struttura dei componenti non

collassata con il maggiore spread globale. Le immagini decodate da campioni latenti GMJE (Fig. 10c) sono visivamente più nitide e semanticamente diverse rispetto alle chimere sfuocate di SimCLR (Fig. 10a) e ai digit sovra-lisciati di GJE (Fig. 10b). Questo pattern di risultati — latent samples ben distribuiti sui cluster semantici, immagini decodate riconoscibili e diverse — è la conferma empirica diretta che il manifold latente GMJE è intrinsecamente generativo, non soltanto discriminativo.

III.8.A — Considerazioni sull'implementazione pratica di I-JEPA

La descrizione formale di I-JEPA nei paragrafi precedenti rischia di oscurare alcune considerazioni pratiche di implementazione che hanno impatto significativo sulla qualità delle rappresentazioni apprese. Questa sezione documenta le scelte implementative chiave e la loro motivazione.

La scelta dell'ottimizzatore. I-JEPA usa AdamW con weight decay $\lambda = 0.05$ e learning rate massimo $lr_max = 1.5 \times 10^{-3}$ con schedule cosineano. La scelta di AdamW rispetto ad Adam standard è motivata dalla necessità di una regolarizzazione robusta per prevenire il sovraddattamento su dataset di dimensioni moderate. Il weight decay agisce come una forma di regolarizzazione L2 sui pesi ma con l'importante distinzione che viene applicato ai pesi direttamente, non al gradiente — una distinzione che diventa rilevante quando si usa il momentum dell'ottimizzatore.

Il warm-up del learning rate. Il training di I-JEPA inizia con un warm-up lineare del learning rate per le prime 15 epoche, partendo da $lr = 0$. Questo warm-up è essenziale per prevenire instabilità nelle prime fasi del training quando sia l'encoder che il target encoder sono inizializzati casualmente e producono target privi di significato. Il warm-up dà al sistema il tempo di sviluppare rappresentazioni parzialmente strutturate prima di applicare il pieno gradiente, evitando che le fasi iniziali di training portino i parametri in regioni dello spazio dei pesi da cui è difficile uscire.

Il batch size e la sua relazione con il collasso. La probabilità di collasso delle rappresentazioni in I-JEPA aumenta con batch size molto piccoli. Con batch size = 32, il sistema è suscettibile di collasso parziale nelle prime epoche perché il segnale di errore predittivo è troppo rumoroso per discriminare tra diverse configurazioni dell'encoder. Il batch size standard di 2048 su ViT-L con 16 GPU fornisce un segnale di gradiente sufficientemente stabile da mantenere la dinamica lontana dalla soluzione degenere. Questa dipendenza dal batch size è un aspetto della teoria non completamente formalizzato: esistono argomenti informali basati sulla varianza del gradiente, ma una caratterizzazione precisa del batch size minimo necessario per stabilità rimane un problema aperto.

La periodicità degli aggiornamenti EMA. Nella maggior parte delle implementazioni, il target encoder viene aggiornato con EMA ad ogni step di ottimizzazione — non ad ogni epoca. Questo aggiornamento frequente è

essenziale per mantenere il target encoder sempre abbastanza vicino al context encoder da fornire target significativi, pur sufficientemente ritardato da prevenire il collasso. Un aggiornamento meno frequente (ogni N step) produrrebbe un target encoder troppo stabile che non riflette i progressi recenti del context encoder, rendendo il compito predittivo triviale nelle fasi avanzate del training.

III.9 — Limitazioni strutturali e direzioni di miglioramento

Una trattazione onesta di I-JEPA non può esimersi dall'identificare le limitazioni strutturali dell'approccio, che sono di natura sia teorica che pratica.

La prima limitazione è la dipendenza critica dalla qualità dell'encoder target nella fase iniziale del training. Poiché l'aggiornamento EMA parte dalla stessa inizializzazione casuale del contesto encoder, nelle primissime fasi del training sia il contesto encoder che il target encoder producono rappresentazioni casuali e sostanzialmente prive di informazione. Il predictor, di fronte a target casuali, apprende a predire rumore — un comportamento inutile. La rete supera questa fase perché il target encoder si stabilizza prima del contesto encoder (grazie all'EMA), ma la dinamica delle prime epoche di training è fragile e sensibile agli iperparametri di inizializzazione.

La seconda limitazione riguarda la scalabilità del mascheramento a blocchi. Con risoluzioni di immagine molto elevate — 512×512 o superiore — i blocchi target con area del 15-20% dell'immagine diventano molto grandi in senso assoluto, e predire la rappresentazione di regioni così estese diventa un compito di difficoltà quasi insormontabile che può portare alla degenerazione del training. Sono necessarie strategie di mascheramento adattive che controllino la difficoltà assoluta del compito e non solo quella relativa rispetto alla dimensione dell'immagine.

La terza limitazione è l'assenza di un meccanismo esplicito per modellare la struttura causale del mondo. I-JEPA apprende correlazioni tra regioni nell'immagine, ma non distingue tra correlazioni causali (questa regione è un braccio perché l'immagine raffigura una persona) e correlazioni spurie (questa regione ha una certa texture perché le immagini di questa classe tendono ad avere quella texture come artefatto del dataset). Questa limitazione non è specifica di I-JEPA ma è comune a tutti i metodi di apprendimento autosupervisionato che non incorporano interventi causali nella loro struttura di addestramento.

La quarta limitazione, forse la più fondamentale in prospettiva AGI, è che I-JEPA è un modello puramente passivo: impara a predire stati del mondo ma non interagisce con esso, non esegue azioni, non riceve feedback sulle conseguenze delle proprie predizioni nel mondo reale. Superare questa limitazione richiede di integrare I-JEPA in un sistema più ampio che comprenda un modello del mondo per la pianificazione — ovvero un World Model nel senso dei capitoli successivi.

III.X — Gaussian Joint Embeddings: verso una JEPA probabilistica

Le architetture analizzate nelle sezioni precedenti — da I-JEPA a LeJEPA fino alla Semantic Tube Prediction — condividono un elemento strutturale comune: il

compito predittivo è formulato come regressione deterministica. Il predittore g(z_c) produce un singolo vettore $\hat{z}_t$ che deve approssimare la rappresentazione target z_t minimizzando la distanza quadratica $\|\hat{z}_t - sg(\bar{E}_\theta(x_T))\|^2$. Come dimostrato nella sezione L.2 del paper Gaussian Joint Embeddings (Huang, 2026), minimizzare l'MSE converge al predittore di Bayes ottimale, che è la media condizionale E[z_t | z_c]. In ambiti genuinamente multimodali — dove un contesto z_c è compatibile con più target semanticamente distinti — questo predittore produce la media delle modalità, collocando le predizioni in regioni dello spazio latente non corrispondenti ad alcun campione reale. Il framework dei Gaussian Joint Embeddings (GJE), introdotto da Huang (2026), risponde a questa diagnosi sostituendo la predizione deterministica con la modellazione generativa della densità congiunta p(z_c, z_t).

La differenza concettuale rispetto a I-JEPA è illustrata in modo nitido dalla struttura dell'obiettivo. In I-JEPA si ottimizza: $\mathcal{L}_JEPA = \mathbb{E}\ [\|P_\varphi(E_\theta(x_C), M) - sg(\bar{E}_\theta(x_T))\|^2]$, che è una stima del rischio condizionale p(z_t | z_c) senza esplicita modellazione della densità marginale p(z_c). In GJE si ottimizza invece la verosimiglianza negativa della distribuzione congiunta: $\mathcal{L}_GJE = -\mathbb{E}\ [\log p(E_\theta(x_c), \bar{E}_\theta'(x_t))]$, che per la regola della catena si fattorizza come $-\log p(z_c, z_t) = -\log p(z_t \mid z_c) - \log p(z_c)$. Questo secondo termine, assente in I-JEPA, è la densità marginale del contesto e svolge una funzione di regolarizzazione geometrica: forza le rappresentazioni del contesto a mantenere volume non degenere nello spazio latente, riducendo la dipendenza dalle asimmetrie architetturali (EMA, stop-gradient) per prevenire il collasso.

III.X.1 — La formulazione duale: Gaussian Process Joint Embeddings

La formulazione duale (GPJE) opera nello spazio dei campioni, chiedendo quanto siano simili le immagini tra loro piuttosto che come le feature individuali co-varino. Sia $k\phi$ un kernel definito-positivo (e.g. RBF) sui vettori di contesto. Per un batch di N coppie (Z_c, Z_t), si costruisce la matrice di Gram N×N: $K_cc[i,j] = k\phi(z_c,i, z_c,j)$. Assumendo che ogni canale dimensionale di Z_t sia un Gaussian Process indipendente con covarianza K_cc, la log-verosimiglianza marginale (NLL condizionale) è:

$$\mathcal{L}_{Dual\text{-}GJE}(\theta, \theta', \phi) = (1/2)\ Tr(Z_t^T\ K_cc^{-1}\ Z_t) + (d_t/2)\ \log\ |K_cc|$$

Il termine $Tr(Z_t^T\ K_cc^{-1}\ Z_t)$ è la data-fit: penalizza la distanza di Mahalanobis tra le rappresentazioni target e lo spazio del contesto. Il termine $(d_t/2)\ \log|K_cc|$ è la penalizzazione di complessità (Occam's razor): promuove smoothness del context manifold. La predizione inferenziale per un nuovo contesto z^*_c avviene attraverso la distribuzione condizionale gaussiana esatta: $\mu_{t|c} = k^*_c\ K_cc^{-1}\ Z_t$, con $k^*_c \in \mathbb{R}^{1\times N}$ il vettore di cross-similarità. Il risultato è una media pesata non-parametrica sull'intero training bank Z_t — una interpolazione globale nello spazio semantico che non richiede un predittore neurale separato.

Questa formulazione rivela un fatto strutturale importante: la Dual-GJE è intrinsecamente asimmetrica. Ottimizza la verosimiglianza condizionale p(Z_t | Z_c), penalizzando il volume del context space (log|K_cc|) ma non imponendo alcun regolarizzatore espansivo sullo spazio target. Come dimostrato nell'Appendice E di Huang (2026), ottimizzare questa perdita simmetricamente produce il collasso catastrofico: il target encoder converge verso Z_t → 0 per minimizzare il termine quadratico, dopodiché il context encoder collassa su una costante per minimizzare il log-determinante. Il meccanismo EMA del target encoder in I-JEPA è la soluzione ingegneristica a questo problema strutturale, ora derivabile formalmente dalla matematica del GP.

III.X.2 — La formulazione primale: covariance feature-space

La formulazione primale opera nello spazio delle feature, costruendo la matrice di covarianza d×d anziché la Gram N×N. Sia Z = [Z_c, Z_t] ∈ ℝ^{N×d} la concatenazione degli embedding. Si assume un modello generativo gaussiano a media zero: p(z) = $\mathcal{N}$(0, C_{joint}), con C_{joint} = (1/N) Z^T Z. La log-verosimiglianza empirica negativa (NLL media) si scrive:

$$\mathcal{L}_{\text{Primal-GJE}}(\theta, \theta') = (1/2N)\ \Sigma_i\ z_i^T\ C_{\text{joint}}^{-1}\ z_i + (1/2)\ \log\ |C_{\text{joint}}|$$

La partizione a blocchi C_{joint} = [[C_cc, C_ct],[C_tc, C_tt]] cattura sia le auto-covarianze (C_cc, C_tt) sia la cross-covarianza C_ct che codifica l'informazione mutua tra contesto e target. Dalla distribuzione congiunta gaussiana si deriva analiticamente la distribuzione condizionale p(z*_t | z*_c) = $\mathcal{N}$(μ_{t|c}, Σ_{t|c}) con parametri in forma chiusa: μ_{t|c} = C_tc C_cc^{-1} z*_c e Σ_{t|c} = C_tt − C_tc C_cc^{-1} C_ct. Il predittore ottimale è dunque una proiezione lineare derivata dalle covarianze globali, senza necessità di addestrare un predittore neurale separato.

L'equivalenza primale-duale (Appendice J di Huang, 2026) stabilisce che con kernel lineare (K_cc = Z_c Z_c^T), la Dual-GJE collassa algebricamente nella Primal-GJE: le due formulazioni ottimizzano esattamente la stessa struttura di covarianza d×d. Questa equivalenza matematica dimostra che senza approssimazioni kernelizzate non lineari (RBF, RFF), il modello dual e primale condividono la medesima limitazione strutturale: sono intrinsecamente unimodali. Non possono instradare le predizioni verso rami semantici distinti quando il target ha distribuzione multimodale.

III.X.3 — Il Mahalanobis Trace Trap: dimostrazione formale del collasso algebrico

L'implementazione diretta della Primal-GJE nella pratica dell'apprendimento profondo incontra un ostacolo algebrico fondamentale che Huang (2026) denomina il Mahalanobis Trace Trap. Il problema emerge ogni volta che si tenta di ottimizzare la verosimiglianza congiunta empirica usando la covarianza calcolata dinamicamente dal batch corrente. La dimostrazione è elementare ma le sue conseguenze sono catastrofiche.

Sia C = (1/N) Σ_i z_i z_i^T la covarianza empirica calcolata dallo stesso batch Z di N vettori. Il termine di data-fit nella NLL è: L_Maha = (1/N) Σ_i z_i^T C^{-1} z_i. Poiché ogni scalare quadratico è uguale alla propria traccia, e per la proprietà ciclica Tr(AB) = Tr(BA):

L_Maha = (1/N) Σ_i Tr(C^{-1} z_i z_i^T) = Tr(C^{-1} · (1/N) Σ_i z_i z_i^T) = Tr(C^{-1} C) = Tr(I_d) = d

L'intera distanza di Mahalanobis si cancella algebricamente, collassando in una costante scalare d indipendente dai parametri della rete. Il gradiente ∂L_Maha/∂θ = 0: il termine di data-fit non produce alcun segnale di apprendimento. L'ottimizzatore rimane con sola la penalizzazione del volume +(1/2) log|C_{joint}|, che viene minimizzata spingendo tutte le rappresentazioni verso l'origine, causando il collasso dimensionale catastrofico.

Il Trap colpisce anche nel caso multimodale (GMM): per il componente k con responsabilità γ_ik e covarianza empirica Σ_k = (1/N_k) Σ_i γ_ik(z_i − μ_k)(z_i − μ_k)^T, il termine di data-fit pesato vale Σ_i γ_ik (z_i − μ_k)^T Σ_k^{-1} (z_i − μ_k) = N_k · Tr(Σ_k^{-1} Σ_k) = N_k · d. Ogni componente del GMM subisce lo stesso annullamento: nessun gradiente attrattivo sopravvive. Questa è la ragione strutturale per cui le tre varianti architetturali di GMJE (prototipi globali, MDN, SMC) non sono mere scelte di design ma rimedi necessari al Trap: ciascuna disaccoppia i parametri di covarianza dalla covarianza empirica del batch corrente.

III.X.4 — Gaussian Mixture Joint Embeddings: il framework multimodale

Per superare la limitazione unimodale di GJE, il framework Gaussian Mixture Joint Embeddings (GMJE) sostituisce la singola gaussiana congiunta con una mistura di K componenti. La densità congiunta diventa: p(z_c, z_t) = Σ_{k=1}^K π_k $\mathcal{N}$([z_c; z_t] | [μ_c,k; μ_t,k], Σ_k). Per il teorema di approssimazione universale delle GMM (Teorema 1 di Huang, 2026, che estende Li e Barron, 1999), qualsiasi densità continua su ℝ^d può essere approssimata in norma L1 con accuratezza arbitraria da una mistura gaussiana con numero finito K di componenti.

Dalla densità congiunta si derivano analiticamente le distribuzioni marginale e condizionale. La marginale p(z_t) = Σ_k π_k $\mathcal{N}$(z_t | μ_t,k, Σ_tt,k) rimane una GMM chiusa. La condizionale p(z_t | z_c) = Σ_k γ_k(z_c) $\mathcal{N}$(z_t | μ_{t|c,k}, Σ_{t|c,k}) è anch'essa una GMM, con pesi data-dipendenti γ_k(z_c) = π_k $\mathcal{N}$(z_c | μ_c,k, Σ_cc,k) / Σ_j π_j $\mathcal{N}$(z_c | μ_c,j, Σ_cc,j) e parametri condizionali per componente: μ_{t|c,k} = μ_t,k + Σ_tc,k Σ_cc,k^{-1}(z_c − μ_c,k), Σ_{t|c,k} = Σ_tt,k − Σ_tc,k Σ_cc,k^{-1} Σ_ct,k. Il vettore di contesto z_c svolge dunque il ruolo di router probabilistico: valuta la sua verosimiglianza marginale sotto ciascun componente e pesa le predizioni di conseguenza. Architetture come JEPA con predittore MLP approssimano questa struttura in modo implicito; GMJE la rende esplicita e ottimizzabile con garanzie teoriche.

La variante prototipica (GMJE-Proto, sezione 4.3 di Huang, 2026) mantiene K prototipi globali apprendibili μ_k = [μ_c,k; μ_t,k] e una covarianza condivisa Σ,

ottimizzando la NLL della mistura attraverso il Log-Sum-Exp (LSE) numericamente stabile: L_Proto = −log Σ_k exp(log π_k − (1/2)(Z − μ_k)^T Σ^{-1}(Z − μ_k) − (1/2)log|Σ|). Il gradiente dell'LSE rispetto ai prototipi produce esattamente le probabilità softmax, che corrispondono alle responsabilità posteriori γ_k(Z): la maggior parte del segnale di aggiornamento fluisce verso il prototipo dominante, realizzando un clustering differenziabile end-to-end. Questa architettura bypassa il Mahalanobis Trace Trap perché Σ è un parametro apprendibile globale, non la covarianza empirica del batch.

III.X.5 — GMJE-MDN e GMJE-GNG: istanze parametriche dinamiche

La variante GMJE-MDN (Mixture Density Network, sezione 4.4 di Huang, 2026) estende GMJE-Proto permettendo che i parametri della mistura siano funzioni dinamiche del contesto: μ_k(z_c), Σ_k(z_c), π_k(z_c). Questa flessibilità permette al modello di adattare la forma della distribuzione condizionale all'ambiguità epistemica specifica di ogni input. Il rischio di identity collapse — il predittore che osserva z_t e impara banalmente μ_k(Z) = Z riducendo la varianza a zero — è prevenuto attraverso un bottleneck informativo: la rete parametrica osserva solo z_c, non [z_c; z_t]. Questo vincolo architetturale, derivato formalmente nell'Appendice M, è l'analogo probabilistico del predictor ristretto in I-JEPA: limita la capacità del predittore per forzare l'encoder a produrre rappresentazioni intrinscamente predittive. L'obiettivo completo fattorizza come $\mathcal{L}_{MDN} = -\log p(z_c) - \log p(z_t \mid z_c)$, dove il termine marginale (−log p(z_c)) impedisce il collasso del context encoder.

La variante GMJE-GNG (Growing Neural Gas, sezione 4.5) risolve il problema della scelta di K ricorrendo a un algoritmo di scoperta topologica adattiva. GNG rappresenta i prototipi come nodi di un grafo dinamico: partendo da due nodi inizializzati casualmente, spawn nuovi prototipi nelle regioni con errore di quantizzazione elevato, e sfronta connessioni obsolete attraverso un meccanismo di invecchiamento degli archi. Per ogni embedding Z = [z_c; z_t] campionato, si identificano i due prototipi più vicini s1 e s2, si aggiorna s1 con tasso $\varepsilon_b \gg \varepsilon_n$ e i suoi vicini topologici con tasso ε_n, si accumula errore di quantizzazione Es1 ← Es1 + ‖Z − μ_s1‖², e ogni λ iterazioni si inserisce un nuovo nodo tra u (massimo errore) e il suo vicino v con massimo errore: μ_w = (μ_u + μ_v)/2. L'algoritmo scopre organicamente sia il numero di modalità K sia la topologia del manifold semantico congiunto.

III.X.6 — Il teorema dell'InfoNCE bridge: contrastive learning come caso degenere di GMJE

Uno dei risultati teorici più significativi di Huang (2026) è la dimostrazione formale che l'apprendimento contrastivo standard — nelle sue formulazioni InfoNCE, SimCLR, MoCo — è un caso limite non-parametrico e degenere del framework GMJE. La connessione chiarisce i limiti strutturali dei metodi contrastivi e spiega perché essi funzionano entro certi confini e falliscono fuori da essi.

La derivazione procede in tre passi. Primo: si adotta la formulazione DaM-GMJE (Data as Mode), in cui ogni coppia di embedding storici (z_c^(m), z_t^(m)) dalla memory bank di dimensione M è trattata come un componente equiprobabile della mistura: p(z_c, z_t) = (1/M) Σ_m $\mathcal{N}$([z_c; z_t] | [z_c^(m); z_t^(m)], Σ). Secondo: si restringe la struttura della covarianza Σ a una matrice block-diagonale isotropica Σ = diag(τI, τI), imponendo che la cross-covarianza tra contesto e target sia esattamente zero: Σ_ct = 0. Terzo: si applica la normalizzazione L2 agli embedding e si prende il limite di bassa temperatura (small τ).

Sotto queste tre assunzioni, la NLL condizionale −log p(z_t | z_c) si semplifica algebricamente nella perdita InfoNCE: −log [exp(z_c^T z_t / τ) / Σ_m exp(z_c^T z_t^(m) / τ)]. Questa è esattamente l'obiettivo di MoCo, SimCLR e altri metodi contrastivi. Il risultato ha implicazioni epistemiche profonde. L'assunzione Σ_ct = 0 — che la cross-covarianza tra spazio contesto e spazio target sia zero — implica che le due rappresentazioni appartengano a universi statisticamente disconnessi. Per allinearle nonostante questa disconnessione strutturale, i metodi contrastivi sono costretti a imporre la condivisione dei pesi (Siamese) e a massimizzare la similarità dei dot-product — un workaround architetturale che rimpiazza l'informazione mutua genuina con l'invarianza alle augmentation. Inoltre, la formulazione KDE non-parametrica (ogni campione = un modo) soffre del class collision problem: coppie semanticamente identiche vengono trattate come campioni negativi, sprecando capacità rappresentativa per separare classi già separabili.

Questa gerarchia teorica — GMJE (generale) → DaM-GMJE (non-parametrico) → InfoNCE (degenere isotropico) — è il complemento al framework LeJEPA: mentre LeJEPA deriva la distribuzione ottimale degli embedding dal basso verso l'alto (dal rischio di probing alle proprietà distribuzionali), GMJE unifica le architetture SSL dall'alto verso il basso, mostrando che tutti i metodi esistenti istanziano approssimazioni di un medesimo principio generativo.

III.X.7 — SMC-GMJE: il particle filter come memory bank ottimale

Il framework DaM-GMJE evidenzia un limite dell'implementazione standard dei metodi contrastivi: la memory bank FIFO (First-In-First-Out) di MoCo impone un prior uniforme costante π_m = 1/M su tutti i componenti, indipendentemente dalla loro utilità informativa. Campioni rarissimi o hard negatives vengono espulsi dalla coda semplicemente perché temporalmente vecchi, mentre stream di campioni easy e ridondanti rimangono nella bank. SMC-GMJE (sezione 4.7 di Huang, 2026) sostituisce questo meccanismo deterministico con un particle filter probabilistico che mantiene pesi di importanza dinamici W^(m).

Il ciclo SMC si articola in due passi. Nel passo di importance weighting, dati i nuovi query Z_query del batch corrente, si aggiorna il peso non normalizzato di ciascuna particella m attraverso la sua verosimiglianza media sotto i query: $\tilde{W}$^(m) = W^(m)_{prev} × (1/B) Σ_{z_c ∈ Z_query} exp(z_c^T z_t^(m) / τ). I pesi vengono poi normalizzati. Nel passo di Sequential Importance Resampling (SIR), si forma un pool combinato M+B (bank + batch corrente), si campionano M

particelle con rimpiazzo secondo i pesi posteriori W^(m), e si resettano tutti i pesi a 1/M. L'equivalenza formale con la memory bank pesata è garantita: dopo il SIR, la InfoNCE standard sull'insieme ricampionato è algebricamente identica alla SMC-weighted InfoNCE − log[exp(z_c^T z_t/τ) / Σ_m W^(m) exp(z_c^T z_t^(m)/τ)].

Sperimentalmente (sezione 5.2.1 di Huang, 2026), su CIFAR-10 con ResNet-18 e memory bank ridotta a M=256, SMC-GMJE raggiunge 81.99% di linear probing accuracy contro il 78.08% di MoCo v2 FIFO dopo 1000 epoche — un guadagno di +3.91% con la stessa capacità di memoria. L'Effective Sample Size (ESS = $1/\Sigma(W^{(m)})^2$) del filtro rivela un fenomeno significativo: durante il warmup (prima di epoch 500) ESS rimane alto (distribuzione quasi uniforme), ma quando i cluster semantici cominciano a formarsi, ESS si abbassa bruscamente mentre le hard negatives accumulano pesi crescenti e vengono clonate durante il SIR. Il filtro SMC trasforma automaticamente la memory bank in un magazzino di hard negatives senza richiedere alcuna euristica esterna di mining.

Questo risultato si integra con le stime di compute del Capitolo VI: SMC-GMJE riduce la dimensione minima della memory bank necessaria per una data qualità di rappresentazione, aggiungendo un ulteriore vettore di efficienza computazionale oltre ai fattori ϱ_STP ≈ 16 di Semantic Tube Prediction e alla riduzione dell'overhead di selezione del modello di LeJEPA.

III.X.8 — Generative GMJE: il campionamento incondizionato come proprietà emergente

Una conseguenza strutturale del framework GMJE che non ha analogo in I-JEPA, LeJEPA o nei metodi contrastivi è la disponibilità di una densità generativa p(z_t) in forma chiusa. Poiché GMJE ottimizza la NLL della densità congiunta p(z_c, z_t), la marginalizzazione analitica su z_c produce automaticamente $p(z_t) = \Sigma_k \pi_k \mathcal{N}(z_t \mid \mu_{t,k}, \Sigma_{tt,k})$ — una GMM continua sullo spazio delle rappresentazioni target. Questo permette il campionamento incondizionato di nuovi embedding sintetici senza accesso ad alcun dato reale: (1) si campiona un indice $k \sim \text{Categorical}(\pi_1,\ldots,\pi_K)$, (2) si campiona $z_t \sim \mathcal{N}(\mu_{t,k}, \Sigma_{tt,k})$ via decomposizione di Cholesky.

I risultati sperimentali su MNIST (sezione 5.3 di Huang, 2026) illustrano il contrasto con le alternative. Un post-hoc GMM fittato su embedding SimCLR produce campioni latenti nelle regioni inter-cluster (zone a bassa densità), perché SimCLR non è addestrato come modello generativo e la distribuzione dello spazio latente non è controllata. La GJE unimodale sparge campioni su tutto lo spazio senza struttura semantica (Avg. $\log|\Sigma_k| = -636.15$, prototype spread = N/A). Il GMJE parametrico (K=50 componenti) produce campioni concentrati nei cluster semantici (Avg. $\log|\Sigma_k| = -621.58$, prototype spread = 1.43), con immagini decodate nitide e diverse. Questa capacità generativa ha rilevanza diretta per il framework AMI: il Perception Module GMJE può fungere da generatore latente per continual learning (replay generativo senza storage dei dati originali), data

augmentation latente, e inizializzazione del planning in regioni semanticamente dense.

III.X.9 — Connessione con LeJEPA: chiusura del cerchio teorico

Il framework GJE/GMJE e il framework LeJEPA (sezione III.5.2) affrontano lo stesso problema da direzioni complementari, e la loro connessione chiude un cerchio teorico che nessuno dei due apre completamente da solo. LeJEPA risponde alla domanda normativa: quale distribuzione marginale p(z) deve avere un encoder SSL per essere ottimale rispetto a qualsiasi compito downstream? Risposta: la gaussiana isotropica $\mathcal{N}(0, (\varkappa/K)I)$. GJE risponde alla domanda strutturale: quale obiettivo di addestramento produce distribuzioni latenti con la geometria corretta e la predizione non degenere? Risposta: la verosimiglianza congiunta $-\log p(z_c, z_t)$ con modellazione esplicita delle covarianze.

La connessione è precisa: il termine marginale $-\log p(z_c)$ nella fattorizzazione della GJE joint NLL ottimizza la distribuzione del context encoder verso la gaussiana a volume massimo. Con la regolarizzazione entropica (maximization di $(1/2)\log|C_cc|$ invece della minimizzazione) che sostituisce il Mahalanobis morto, l'obiettivo GJE spinge $p(z_c)$ verso configurazioni isotropiche — la stessa distribuzione che LeJEPA derivia ottimale tramite il teorema di Cramér-Rao. SIGReg di LeJEPA può dunque essere vista come l'implementazione efficiente del termine marginale di GJE: mentre GJE usa la log-det della covarianza empirica, SIGReg usa proiezioni stocastiche e test di Epps-Pulley, ottenendo la stessa pressione distribuzionale con gradienti uniformemente limitati e complessità lineare.

La triangolazione teorica GJE–LeJEPA–STP è quindi completa: GJE fonda la predizione latente su una densità congiunta esplicita (struttura della distribuzione di coppia), LeJEPA caratterizza l'ottimale distribuzione marginale per il probing downstream (struttura delle singole rappresentazioni), STP regolarizza la geometria delle traiettorie temporali (struttura dell'evoluzione delle rappresentazioni nel tempo). Questi tre vincoli — covarianza di coppia, isotropia marginale, linearità delle traiettorie — sono ortogonali e complementari: il loro impianto congiunto in un sistema AMI rappresenta la formulazione teoricamente più completa disponibile alla data di questa monografia per un Perception Module probabilisticamente fondato.

III.Y — Rectified LpJEPA: Rappresentazioni Sparse e a Massima Entropia

La triangolazione teorica GJE–LeJEPA–STP analizzata nelle sezioni precedenti condivide un presupposto implicito che merita di essere esplicitato e messo in discussione: l'assunzione che la distribuzione ottimale degli embedding per qualsiasi compito downstream sia la gaussiana isotropica, distribuita su tutto lo spazio latente. Questo presupposto — dimostrato rigoroso dal Teorema 1 di LeJEPA per il rischio di probing lineare e non lineare — produce rappresentazioni necessariamente dense: ogni coordinata latente ha varianza non nulla e nessun valore è sistematicamente zero. Eppure la sparsità — la proprietà per cui la maggior parte delle coordinate di una rappresentazione è esattamente zero —

emerge ripetutamente come caratteristica fondamentale delle codifiche efficienti in neuroscienze, elaborazione del segnale e apprendimento profondo. Il paper "Rectified LpJEPA" di Kuang, Dagade, Rudner, Balestriero e LeCun (arXiv:2602.01456, febbraio 2026) risponde a questa tensione dimostrando che sparsità e massima entropia non sono obiettivi in conflitto: esiste una famiglia di distribuzioni che massimizza l'entropia sotto vincoli di norma ℓp e induce simultaneamente sparsità ℓ0 attraverso la rettificazione.

III.Y.1 — La distribuzione Gaussiana Generalizzata Rettificata (RGG)

Il punto di partenza è la Gaussiana Generalizzata GN_p(μ, σ), la cui densità è proporzionale a exp(−|x − μ|^p / pσ^p). Per p = 2 si ottiene la gaussiana standard; per p = 1 la distribuzione di Laplace. La Gaussiana Generalizzata Troncata TGN_p(μ, σ) restringe il supporto a (0, +∞) ed è dimostrata essere la distribuzione a massima entropia differenziale soggetta al vincolo E[‖x‖_p^p] = dσ^p — un risultato che generalizza il noto fatto che la gaussiana massimizza l'entropia sotto vincoli di varianza. La Gaussiana Generalizzata Rettificata RGN_p(μ, σ) è invece una mistura tra una misura di Dirac in zero e la distribuzione troncata: si ottiene applicando la ReLU a una variabile aleatoria GN_p(μ, σ), ovvero x ~ ReLU(GN_p(μ, σ)). La sua densità è mista — discreta in 0 con massa Φ_GNp(−μ/σ) e assolutamente continua su (0, +∞) — rendendo la differenziale usuale inapplicabile. Si ricorre alla dimensione di informazione di Rényi d(ξ) = D · Φ_GNp(μ/σ), che misura la quota di coordinate effettivamente continue della distribuzione. Il Teorema 3.6 del paper dimostra che la d(ξ)-entropia della RGG preserva la proprietà di massima entropia dell'originale, riscalata per la dimensione di informazione di Rényi. In particolare il valore atteso della norma ℓ0 è controllato analiticamente: E[‖x‖_0] = d · Φ_GNp(μ/σ), dove Φ_GNp è la CDF della gaussiana generalizzata standard. Modulando μ — lo shift di media — si controlla quindi la frazione attesa di coordinate non nulle con precisione arbitraria, passando dal regime denso (μ → +∞) al regime sparse estremo (μ → −∞).

III.Y.2 — RDMReg: la necessità del test a due campioni

SIGReg di LeJEPA sfrutta una proprietà cruciale della gaussiana isotropica: essa è chiusa rispetto alle combinazioni lineari, ovvero qualsiasi proiezione lineare $c^T z$ di una gaussiana isotropica rimane gaussiana. Ciò consente di confrontare le proiezioni univariate con la gaussiana standard usando un test parametrico (Epps-Pulley). La RGG non gode di questa proprietà: come illustrato nella Figura 1b del paper, le marginali proiettate di una RGG lungo direzioni diverse non appartengono più alla famiglia RGG. Questo è il motivo per cui la Rectified Distribution Matching Regularization (RDMReg) deve adottare un test a due campioni non parametrico: si campionano y ~ ∏_{i=1}^D RGN_p(μ, σ) e si confrontano le proiezioni univariate $c^T z$ e $c^T y$ attraverso la distanza di Wasserstein affettata (sliced 2-Wasserstein). La perdita è definita come L(P_{$c^T z$} ‖ P_{$c^T y$}) = (1/B) ‖(Zc)↑ − (Yc)↑‖², dove la freccia indica l'ordinamento in senso crescente. La perdita completa di Rectified LpJEPA è quindi la somma di un termine di invarianza tra viste e della RDMReg applicata agli embedding rettificati z^+ = ReLU(f_θ(x)):

$\min_\theta \mathbb{E}_{z,z'}[\|z - z'\|^2] + \mathbb{E}_c[L(P_{c^T z} \parallel P_{c^T y})] + \mathbb{E}_c[L(P_{c^T z'} \parallel P_{c^T y})]$

dove c è campionato uniformemente dalla sfera $\ell 2$, e y è campionato dalla distribuzione target $RGN_p(\mu, \sigma)$ prodotto su tutte le D coordinate. Il Lemma I.1 dimostra che la minimizzazione di RDMReg recupera implicitamente una forma di Non-Negative VCReg: se le proiezioni sugli autovettori della covarianza empirica sono pari a quelle della RGG, allora la covarianza degli embedding rettificati è $\gamma \cdot I_D$, ovvero isotropica. In questo senso Rectified LpJEPA generalizza strettamente LeJEPA: ponendo $p = 2$, $\mu = 0$, $\sigma = \sigma_{GN}$ e rimuovendo la rettificazione si recupera esattamente l'obiettivo SIGReg con target gaussiano isotropico.

III.Y.3 — Sparsità controllabile e tradeoff sparsità–prestazione

I risultati empirici del paper documentano tre proprietà notevoli di Rectified LpJEPA. Primo, la sparsità è controllabile: le norme $\ell 0$ empiriche degli embedding preaddestrati seguono con alta fedeltà le previsioni teoriche della Proposizione 3.5, al variare di μ e p, su architetture diverse (ResNet, ViT, ConvNeXt). Questo rende Rectified LpJEPA il primo sistema di apprendimento autosupervisionato con garanzie quantitative sul livello di sparsità degli embedding. Secondo, il tradeoff sparsità–prestazione è favorevole: la frontiera di Pareto tra sparsità $\ell 0$ e accuratezza di probing lineare su CIFAR-100 mostra un decadimento lento e regolare, con un crollo netto soltanto quando più del 95% delle coordinate è zero. Terzo, le rappresentazioni apprese sono più indipendenti statisticamente: il criterio HSIC normalizzato risulta sistematicamente inferiore a quello di VICReg e NVICReg, che penalizzano soltanto le dipendenze di secondo ordine, avvicinandosi a codifiche fattoriali genuine. Le prestazioni di linear probing su ImageNet-100 (fino a 85,08% per l'encoder con target $RGN_2(0, \sigma_{GN})$) sono competitive con le baseline dense, dimostrando che l'induzione di sparsità non sacrifica la qualità delle rappresentazioni per compiti di classificazione.

La capacità di sparsità adattiva ai dataset merita attenzione particolare nel contesto AMI. Nelle Figure 12 e 4c del paper si osserva che Rectified LpJEPA esibisce la maggiore variazione di sparsità tra dataset diversi (DTD, Flowers, CIFAR-10, Food, Pets) rispetto a qualsiasi altra baseline. Questa variabilità — che gli autori denominano dataset-adaptive sparsity — suggerisce che la sparsità delle rappresentazioni possa fungere da proxy per distinguere dati in distribuzione da dati fuori distribuzione, una proprietà di diretta rilevanza per il Cost Module di AMI nella sua funzione di quantificazione dell'incertezza epistemica. Un Perception Module che produce embedding di cui si può misurare la sparsità per rilevare l'OOD si avvicina concettualmente alla stima dell'incertezza epistemica discussa nella sezione V.5, con il vantaggio di non richiedere ensemble o inferenza bayesiana approssimata.

III.Y.4 — La famiglia RGG: Definizioni, Proprietà e Campionamento

La trattazione matematica completa della famiglia RGG parte dalla Gaussiana Generalizzata $GN_p(\mu, \sigma)$, la cui densità di probabilità è definita formalmente come $f(x) = (p^{1-1/p} / (2\sigma\Gamma(1/p))) \exp(-|x - \mu|^p / (p\sigma^p))$, dove $\Gamma(\cdot)$ è la

funzione gamma. Per p = 2 si ottiene la gaussiana standard N(μ, σ^2); per p = 1 la distribuzione di Laplace L(μ, σ); per p → ∞ la distribuzione uniforme. La CDF di GN_p(0, 1) valutata in μ/σ determina la proporzione di massa sopra zero e quindi la sparsità attesa della distribuzione rettificata. La Gaussiana Generalizzata Troncata TGN_p(μ, σ, S) restringe il supporto a un sottoinsieme S di misura di Lebesgue positiva; nel caso di interesse S = (0, +∞), con densità proporzionale a exp(−|x − μ|^p / (pσ^p)) · 1_{(0,∞)}(x), normalizzata dalla funzione di partizione Z_S(μ, σ, p) = ∫_0^∞ exp(−|x − μ|^p / (pσ^p)) dx. La Proposizione 3.3 del paper dimostra che la distribuzione a massima entropia differenziale su S = (0, +∞)^d sotto il vincolo E[‖x‖_p^p] = d · d/dλ_1 log Z_S(λ_1) è esattamente la TGN_p multivariata con densità p(x) = (1/Z_S(λ_1)) exp(−‖x‖_p^p / (pσ^p)) · 1_S(x), dove λ_1 = −1/(pσ^p) è il moltiplicatore di Lagrange. Questo generalizza il risultato noto che la gaussiana isotrópica è l'unica distribuzione a massima entropia sotto vincolo di varianza (caso p = 2), estendendolo all'intera famiglia di norme ℓ_p per ogni p ≥ 0.

La Gaussiana Generalizzata Rettificata RGN_p(μ, σ) è formalmente definita come una mistura tra la misura di Dirac δ_0 e la TGN_p troncata al semi-asse positivo, con la seguente densità rispetto alla misura mista ν = λ + δ_0: f(x) = Φ_{GNp(0,1)}(−μ/σ) · 1_{0}(x) + (p^{1−1/p} / (2σΓ(1/p))) exp(−|x − μ|^p / (pσ^p)) · 1_{(0,+∞)}(x). Il peso della massa discreta in zero è esattamente la probabilità che una variabile GN_p(μ, σ) sia non positiva, ovvero Φ_{GNp(0,1)}(−μ/σ). Una variabile RGN_p(μ, σ) si ottiene campionando X ~ GN_p(μ, σ) e poi rettificando: Y = max(0, X) = ReLU(X). L'Algorithm 1 del paper descrive la procedura di simulazione efficiente: si campiona S ~ Unif{−1, +1}, poi G ~ Gamma(1/p, 1), si pone X ← μ + σ · S · (pG)^{1/p}, e infine Y ← max(0, X). La Proposizione B.9 del paper fornisce le formule in forma chiusa per l'aspettativa e la varianza di RGN_p(μ, σ), necessarie per la scelta del parametro di scala σ: E[X] = (1/2)(μ(1 + sgn(μ/σ) P(1/p, |μ/σ|^p/p)) + p^{1/p}σ Γ(2/p, |μ|^p/(pσ^p))/Γ(1/p)), dove P(·, ·) è la funzione gamma regolarizzata inferiore e Γ(·, ·) è la funzione gamma incompleta superiore. La varianza si ottiene per differenza tra il secondo momento e il quadrato del primo: Var(X) = E[X^2] − (E[X])^2.

III.Y.5 — Entropia d(ξ)-dimensionale e Massima Entropia nel Regime Sparso

La RGG è assolutamente continua rispetto alla misura mista ν = λ + δ_0 (Lemma B.6 del paper), non rispetto alla misura di Lebesgue λ. Ciò rende l'entropia differenziale classica inapplicabile: la derivata di Radon-Nikodym dP_X/dλ non esiste. Il Teorema 3.6 del paper ricorre alla dimensione di informazione di Rényi d(ξ) = D · Φ_{GNp(0,1)}(μ/σ), che misura la quota di coordinate effettivamente continue della distribuzione mista, e alla corrispondente d(ξ)-entropia di Rényi (1959): H_{d(ξ_i)}(ξ_i) = Φ_{GNp(0,1)}(μ/σ) · H_1(TGN_p(μ, σ)) + H_0(1_{(0,+∞)}(ξ_i)), dove H_1(·) è l'entropia differenziale della TGN troncata e H_0(·) è l'entropia di Shannon della variabile di Bernoulli che vale 1 con probabilità Φ_{GNp(0,1)}(μ/σ). Per la distribuzione multivariata prodotto si ha H_{d(ξ)}(ξ) = Σ_{i=1}^D H_{d(ξ_i)}(ξ_i) = D · H_{d(ξ_i)}(ξ_i) per

l'indipendenza coordinata. Crucialmente, il Teorema dimostra che la RGG preserva la proprietà di massima entropia della TGN originale: la RGG è la distribuzione mista che massimizza la d(ξ)-entropia soggetta al vincolo E[‖x‖_p^p] = dσ^p e alla struttura di rettificazione. Questa proprietà è ciò che rende la RGG non un compromesso ma la soluzione ottimale al trade-off sparsità-entropia: massima entropia (nel senso opportunamente ridefinito tramite d(ξ)) non è incompatibile con sparsità ℓ_0 controllabile.

Il Lemma F.6 del paper completa il quadro stabilendo che la d(ξ)-entropia di Rényi coincide con l'entropia definita rispetto alla misura mista ν = λ + δ_0, ovvero $H_\nu(Z) = -\int dP_{Z}/d\nu \log(dP_{Z}/d\nu)\, d\nu = d(Z) \cdot H_1(P_{X|(0,+\infty)}) + d(Z) \log(1/d(Z)) + (1-d(Z)) \log(1/(1-d(Z)))$. Questo risultato ha un'implicazione pratica importante: la decomposizione della correlazione totale in termini di entropie marginali e congiunte — $TC(z) = \sum_i H_{d(z_i)}(z_i) - H_{d(z)}(z)$ — continua a valere con la d(ξ)-entropia al posto dell'entropia differenziale, anche per distribuzioni miste non assolutamente continue rispetto a Lebesgue. Ciò significa che le rappresentazioni sparse di Rectified LpJEPA possono essere analizzate con gli stessi strumenti teorici dell'informazione usati per le distribuzioni continue, a patto di adottare la metrica corretta. In pratica, la stima della d(ξ)-entropia per campioni empirici si effettua calcolando $\hat{d}(Z) = (1/B) \sum_i 1_{(0,+\infty)}(z_i)$ (proporzione di attivazioni positive), poi stimando l'entropia differenziale sulle sole attivazioni positive tramite lo stimatore m-spacing di Vasicek (1976): $\hat{H}_1(P_{X|(0,+\infty)}) = (1/(B' - m)) \sum_{i=1}^{B'-m} \log((B'+1)/m \cdot (z_{(i+m)} - z_{(i)}))$, dove $z_{(i)}$ sono i campioni ordinati in senso crescente e m è un iperparametro di spaziatura.

III.Y.6 — Il Recupero di Non-Negative VCReg (Proposizione I.1)

La Proposizione I.1 del paper formalizza il legame tra RDMReg e VICReg, stabilendo un risultato di regolarizzazione implicita di secondo ordine che si ricava come conseguenza della convergenza distributiva verso la RGG. Sia $z \in \mathbb{R}^d$ un vettore di feature con matrice di covarianza $\mathrm{Cov}[z] = \Sigma$, e sia $\Sigma = U\Lambda U^T$ la sua decomposizione spettrale con autovettori $\{u_i\}_{i=1}^d$. Sia $y \sim \prod_{i=1}^d \mathrm{RGN}_p(\mu, \sigma)$ il vettore target e sia $\gamma = \mathrm{Var}[\mathrm{RGN}_p(\mu, \sigma)] \in (0, +\infty)$. Se $u_i^T z =^d u_i^T y$ per ogni $i \in \{1, \ldots, d\}$ — ovvero se le proiezioni unidimensionali lungo tutti gli autovettori della covarianza sono distribuiti come la corrispondente proiezione della RGG target — allora $\Sigma = \gamma \cdot I_d$. La dimostrazione procede osservando che $\mathrm{Var}(u_i^T y) = u_i^T(\gamma I_d)u_i = \gamma$ per ogni u_i unitario, poiché la RGG ha covarianza isotropica γI_d. Per l'assunzione $u_i^T z =^d u_i^T y$ si ha $\mathrm{Var}(u_i^T z) = \mathrm{Var}(u_i^T y) = \gamma$. Poiché u_i è autovettore di Σ, si ha $\mathrm{Var}(u_i^T z) = u_i^T \Sigma u_i = \lambda_i$, donde $\lambda_i = \gamma$ per ogni i, ovvero $\Lambda = \gamma I_d$ e quindi $\Sigma = \gamma I_d$. Sostituendo nella decomposizione $\Sigma = U(\gamma I_d)U^T = \gamma UU^T = \gamma I_d$. Tutti gli elementi fuori diagonale di Σ sono nulli e la covarianza è isotrópica.

Il corollario pratico è che campionando le direzioni di proiezione c_i come autovettori della matrice di covarianza empirica delle feature — invece di direzioni casuali uniformi sulla sfera — si ottiene una convergenza più rapida verso

l'isotropia di secondo ordine, con riduzione sistematica sia della variance loss che della covariance loss nel senso di VICReg. La Figura 14 del paper documenta empiricamente su CIFAR-100 con target RGN_1(μ, σ_GN) che l'incorporazione degli autovettori nelle proiezioni ("Rand + Full Eig") produce una riduzione drastica della covariance loss già nelle prime 200 epoche rispetto alle sole proiezioni casuali ("Rand"), con convergenza finale migliore dell'accuratezza del projector in tutti e tre i regimi di mean shift $\mu \in \{-1, 0, 1\}$ testati. La strategia "Rand + Bottom Eig" — che mescola gli autovettori con i più piccoli autovalori con proiezioni casuali — bilancia meglio il trade-off tra riduzione delle dipendenze di secondo ordine e accuratezza finale, poiché vincolare tutti gli autovettori impone un'isotropia eccessiva nelle fasi iniziali del training quando la distribuzione delle feature non è ancora stabilizzata. Poiché la matrice delle feature $Z \in \mathbb{R}^{B\times D}$ è sempre ottenuta dopo rettificazione ($Z \geq 0$), RDMReg recupera di fatto una forma di Non-Negative VCReg che regolarizza le feature non-negative verso una covarianza isotropica — un analogo del Non-Negative Matrix Factorization nel linguaggio dell'apprendimento autosupervisionato, come discutono gli autori nell'Appendice I.

III.Y.7 — Analisi degli Iperparametri: σ_GN versus σ_RGN

La scelta del parametro di scala σ per la distribuzione target RGN_p(μ, σ) è l'iperparametro più delicato di Rectified LpJEPA, poiché determina sia la varianza della distribuzione prima della rettificazione (Var[GN_p(μ, σ)]) sia quella dopo (Var[RGN_p(μ, σ)]). Il paper Appendice D analizza sistematicamente due strategie. La prima, denotata σ_GN, fissa la varianza della GN_p prima della rettificazione a 1 attraverso $\sigma_GN = \Gamma(1/p)^{1/2} / (p^{1/p} \cdot \Gamma(3/p)^{1/2})$, la formula in forma chiusa derivata dalla Proposizione B.9 con il vincolo Var[GN_p(0, σ_GN)] = 1. Il valore di σ_GN è invariante rispetto a μ (come si osserva nella Figura 8a del paper) e decresce monotonamente con p: $\sigma_GN \approx 1.00$ per $p = 2$ (gaussiana standard), ≈ 0.57 per $p = 0.75$ e ≈ 0.10 per $p = 0.25$. La seconda strategia, denotata σ_RGN, determina σ per rendere unitaria la varianza dopo la rettificazione, Var[RGN_p(μ, σ_RGN)] = 1, attraverso una ricerca per bisezione (Algorithm 2 del paper) sulla formula di varianza della Proposizione B.9. A differenza di σ_GN, il valore di σ_RGN dipende sia da p che da μ (Figura 8b), crescendo significativamente per μ molto negativo: ad esempio $\sigma_RGN \approx 3.40$ per ($p = 2$, $\mu = -3$) contro $\sigma_GN = 1.00$. Questo aumento è necessario perché la rettificazione riduce la varianza drasticamente per μ negativo (fino a Var[RGN_2(−3, σ_GN)] ≈ 0.0002), e σ_RGN compensa questa riduzione espandendo la scala della GN prima della rettificazione.

La Figura 9 del paper confronta teoricamente ed empiricamente le norme ℓ_0 normalizzate sotto le due scelte. La scelta σ_RGN produce rappresentazioni meno sparse di σ_GN a parità di μ e p, perché la varianza unitaria dopo rettificazione richiede una distribuzione pre-rettificazione più allargata, il che aumenta la massa sopra zero. La Figura 10 del paper mostra che, nonostante questa differenza di sparsità, il trade-off sparsità–accuratezza delle due scelte è largamente sovrapposto su CIFAR-100: le frontiere di Pareto di σ_GN e σ_RGN si intersecano senza che

una domini sistematicamente l'altra. Tuttavia, sotto la metrica ℓ_0, la scelta σ_GN con p = 1 (Rectified Laplace) produce il miglior trade-off sparsità–accuratezza nell'insieme, per cui σ_GN viene raccomandata dagli autori come scelta predefinita. La bisezione per trovare σ_RGN converge entro 30 iterazioni con errore sotto 10^{−10} per qualsiasi (p, μ) testato (Figura 11a), rendendola computazionalmente trascurabile nel contesto del training dell'encoder.

III.Y.8 — Indipendenza Statistica, HSIC e Codifiche Fattoriali

Oltre alla sparsità, il paper documenta che Rectified LpJEPA apprende rappresentazioni con minori dipendenze statistiche di ordine superiore rispetto alle baseline. Il Criterio di Hilbert-Schmidt per l'Indipendenza (HSIC) di Gretton et al. (2005) misura la dipendenza non-lineare tra due variabili aleatorie tramite HSIC(X, Y) = (1/(B−1)^2) Tr(KHLH), dove K_{ij} = k(x_i, x_j), L_{ij} = l(y_i, y_j), H = $I - (1/B)11^T$ è la matrice di centraggio, e k, l sono kernel (tipicamente gaussiani). La versione normalizzata nHSIC(X, Y) = HSIC(X, Y) / sqrt(HSIC(X, X) · HSIC(Y, Y)) misura la dipendenza su una scala [0, 1] invariante alla scala. Poiché HSIC cattura dipendenze di qualsiasi ordine — e non solo quelle di secondo ordine catturate dalla matrice di covarianza — serve come misura di indipendenza statistica più completa rispetto alla sola whitening.

La Figura 4b del paper riporta i valori medi dell'nHSIC fuori-diagonale della matrice di dipendenza a coppie per Rectified LpJEPA, LpJEPA, VICReg, NVICReg, CL (SimCLR) e NCL. Rectified LpJEPA ottiene i valori più bassi in assoluto tra i metodi non-contrastivi: sistematicamente inferiori a VICReg e NVICReg (che penalizzano solo le dipendenze di secondo ordine) e comparabili ai metodi contrastivi CL e NCL (che tendono a produrre dipendenze basse grazie alla pressione della separazione dei campioni negativi). Questo risultato è teoricamente previsto: la distribuzione target RGN_p con coordinate i.i.d. ha, per costruzione, dipendenze inter-coordinata nulle a ogni ordine, e RDMReg spinge la distribuzione degli embedding verso questa struttura fattoriale attraverso il test di corrispondenza sulle proiezioni. La convergenza del Cramér–Wold rispetto alla distribuzione RGG garantisce asintoticamente l'indipendenza coordinata completa — non solo la decorrelazione lineare — rendendo Rectified LpJEPA il primo metodo JEPA con garanzia teorica di indipendenza statistica multi-ordine. In pratica, con un numero finito M di proiezioni per step, la convergenza è approssimata, ma la riaccampionamento delle direzioni durante il training produce una copertura cumulativa della sfera che cresce linearmente nel tempo, avvicinando la distribuzione degli embedding alla RGG con alta fedeltà.

III.Y.9 — Risultati Empirici: ImageNet-100, CIFAR-100 e Transfer Learning

La Tabella 1 del paper riporta i risultati di linear probe su ImageNet-100 per Rectified LpJEPA e le baseline dense e sparse. Le varianti di Rectified LpJEPA con target RGN_2(0, σ_GN) e RGN_1(0, σ_GN) raggiungono rispettivamente 85,08% e 84,72% di Acc1 all'encoder, comparabili a LeJEPA (84,80%) e superiori a VICReg (84,18%) e SimCLR (83,44%), pur producendo rappresentazioni sparse (sparsità ℓ_0 rispettivamente 0,730 e 0,694). La configurazione più sparsa testata

— RGN_2(−2.5, σ_GN) con ℓ_0 = 0,022 — mantiene un'accuratezza all'encoder di 82,02%, superiore alla baseline NCL-ReLU (82,58% con ℓ_0 = 0,009). Confrontando Rectified LpJEPA con le baseline sparse NVICReg-ReLU e NCL-ReLU, si osserva che Rectified LpJEPA ottiene il miglior compromesso tra accuratezza e sparsità a parità di regime: per ℓ_0 ≈ 0.7 (regime moderatamente sparso), RGN_1(0, σ_GN) supera NVICReg-ReLU (84,72% vs 84,48%) con sparsità analoga (0,694 vs 0,712). Questo vantaggio è la conseguenza diretta della fondazione teorica: RDMReg impone la distribuzione target RGG ottimale (massima entropia sotto vincoli ℓ_p), mentre NVICReg impone solo la decorrelazione di secondo ordine senza caratterizzare la forma distribuzionale completa.

I risultati di transfer learning in pochi shot (Tabelle 3–8 del paper) confermano questa tendenza. Con 10-shot linear probe su DTD, CIFAR-10, CIFAR-100, Flowers, Food e Pets, RGN_1(0, σ_GN) e RGN_2(0, σ_GN) ottengono un'accuratezza media comparabile o superiore a LeJEPA e VICReg in quasi tutti i dataset, con prestazioni particolarmente robuste in regime 1-shot (Tabella 3). Le varianti molto sparse (μ = −3) degradano più marcatamente in trasferimento, a conferma che la Figura 3c sul trade-off sparsità–accuratezza si generalizza anche ai dataset fuori distribuzione. I risultati con backbone ViT (Tabella 10) replicano la stessa struttura di risultati osservata con ResNet su ImageNet-100, dimostrando che le proprietà di Rectified LpJEPA sono robuste all'architettura dell'encoder. In particolare, RGN_1(0.0, σ_GN) con ViT raggiunge 75,44% di Acc1 all'encoder con sparsità ℓ_0 = 0,772, superando sia VICReg (72,06%) sia SimCLR (74,18%) con sparsità zero — il che suggerisce che la pressione verso la distribuzione RGG produce rappresentazioni intrisecamente più informative, non solo più sparse.

III.Y.10 — Dettagli Implementativi e Connessione con il Framework AMI

L'Appendice L del paper fornisce i dettagli completi dell'implementazione di Rectified LpJEPA su ImageNet-100. L'architettura segue lo schema encoder-projector z = ReLU(f_{θ_2}(f_{θ_1}(x))), dove f_{θ_1} è un encoder (ResNet-50 o ViT-Small) e f_{θ_2} è un MLP a tre layer con dimensione nascosta e di uscita 2048. La ReLU finale è l'elemento che impone la rettificazione esplicita: senza di essa l'architettura non può apprendere rappresentazioni sparse. La loss di training combina un termine di invarianza con peso λ_{sim} = 25.0 e la RDMReg con peso λ_{dist} = 125.0: l'alto peso relativo di RDMReg riflette la necessità di vincolare efficacemente la distribuzione degli embedding verso la RGG target, che richiede più segnale di quello necessario per la sola minimizzazione della perdita di predizione. Il matching distributivo usa la distanza di Wasserstein affettata 2D con 8192 proiezioni casuali per iterazione, un numero sufficientemente grande da garantire una copertura adeguata della sfera in dimensione 2048 senza essere proibitivo computazionalmente (circa 6 ms per forward e backward su una GPU L40S). Il pretraining è condotto per 1000 epoche con ottimizzatore LARS, learning rate 0.0825 per l'encoder con schedule cosineano, batch size 128 su una singola GPU NVIDIA L40S, per una durata complessiva di circa 55 ore.

La connessione di Rectified LpJEPA con il framework AMI si articola su tre livelli complementari. Sul piano del Perception Module, la capacità di produrre embedding sparsi con sparsità ℓ_0 calibrata tramite μ e p arricchisce il repertorio del modulo percettivo: embedding densi (μ = 0) producono alta accuratezza su compiti di classificazione, embedding moderatamente sparsi (μ = −1 o −2) producono codifiche più efficienti in termini di attivazioni non nulle, utili per la trasmissione e l'archiviazione nel framework distribuito di AMI. Sul piano del Cost Module, la dataset-adaptive sparsity — il fatto che le norme ℓ_0 degli embedding varino sistematicamente tra dataset in-distribution e out-of-distribution — offre un segnale di incertezza epistemica senza richiedere ensemble o dropout variazionale. Sul piano del World Model, la combinazione RDMReg + isotropia implicita produce traiettorie di embedding più stabili (come documentato dalla Figura 4b sull'nHSIC), il che si traduce in transizioni di stato latente più regolari nel modello del mondo e quindi in predizioni di orizzonte più lungo con minor compounding error. In prospettiva, la possibilità di modulare il regime di sparsità dell'intero Perception Module con un singolo parametro μ — senza cambiare l'architettura — rende Rectified LpJEPA un candidato naturale per sistemi AMI che operano in condizioni di risorse variabili: più sparsità per l'inferenza su dispositivi edge, meno sparsità per l'addestramento su cluster.

III.10 — L'ipotesi geodetica: traiettorie di token come linee dritte su una varietà semantica

Il principio JEPA nella sua formulazione originale per immagini e video si fonda su una congettura implicita sulla struttura geometrica dello spazio latente: che le rappresentazioni di porzioni diverse di uno stesso input siano mutuamente predittibili attraverso un proiettore di capacità limitata. Questa congettura non riceve giustificazione teorica autonoma nell'architettura I-JEPA, ma viene empiricamente convalidata dal fatto che il sistema converge a rappresentazioni semanticamente ricche. Un lavoro recente di Huang, LeCun e Balestriero (2026) — Semantic Tube Prediction — fornisce per la prima volta una caratterizzazione geometrica esplicita della struttura che lo spazio delle rappresentazioni latenti degli LLM dovrebbe possedere, e costruisce su questa caratterizzazione una loss ausiliaria che migliora l'efficienza dei dati in modo drammatico. Il risultato centrale — che è possibile eguagliare le prestazioni dell'addestramento su un intero dataset usando soltanto 1/16 dei dati — costituisce una violazione diretta del termine dati delle Scaling Laws in stile Chinchilla, e segnala che le leggi empiriche di scala descrivono l'addestramento tipico, non quello ottimale.

Il punto di partenza concettuale è la formalizzazione dell'addestramento di un LLM come sistema dinamico. Sia $x_{\leq t}$ una sequenza di token di lunghezza t, con x_t il t-esimo token e h_t lo stato nascosto corrispondente prodotto dalla rete neurale f al termine del t-esimo passo: $h_t = f(x_{\leq t})$. Si considera il caso della rete convergente, dove le perdite di addestramento sono state minimizzate e dove $\hat{u} \circ f$ denota la composizione delle funzioni del modello a convergenza ($\hat{u}$ è la funzione di unembedding, f è l'encoder). L'equazione di training a convergenza è

$x_{t+1} = \hat{u} \circ \mathcal{f}(x_{\leq t})$, (1) $h_t = \mathcal{f}(x_{\leq t}) + \varepsilon_t$, (2)

dove ε_t è l'errore residuo di unembedding. La chiave del passaggio alla formulazione ODE è che, introducendo un operatore di rimozione del prefisso $\ominus$ definito come $x_{\leq t+1} \ominus x_{\leq t} = v(x_{t+1}, t)$ (dove $v(x, t)$ è il vettore con il token x in posizione t+1 e zeri altrove), la dinamica della sequenza soddisfa

$$x_{\leq t+1} \ominus x_{\leq t} = v(\hat{u} \circ \mathcal{f}(x_{\leq t}), t),$$

che è formalmente equivalente, sotto opportune identificazioni tra embedding discreti e vettori continui in $\mathbb{R}^{T \times d_{model}}$, alla regola di aggiornamento di un'ODE

$$dx_{\leq t} = (\hat{u} \circ \mathcal{f}(x_{\leq t}))\, dt,$$

con condizione iniziale $x_{\leq 0} = [x_1, 0, \ldots, 0]$. Il Teorema di Picard-Lindelöf garantisce che, se $\hat{u} \circ \mathcal{f}(\cdot)$ e le sue derivate parziali rispetto a $x_{\leq t}$ sono continue, questa ODE ammette un'unica soluzione per ogni condizione iniziale. La conseguenza fondamentale è che traiettorie generate da prompt distinti (condizioni iniziali distinte) non possono intersecarsi: l'unicità della soluzione esclude per principio il collasso modale nella generazione, nella misura in cui il modello a convergenza rispetta le condizioni di regolarità.

Alla formulazione ODE si affianca, per il processo di inferenza, una formulazione SDE. In regime di larghezza infinita del modello (limite studiato da Yang e Littwin nel framework dei Tensor Programs), le pre-attivazioni della rete convergono a processi gaussiani. Gli errori ε_t sono quindi approssimativamente gaussiani i.i.d., e il loro accumulo durante la generazione autoregressiva — dove h_{t+1} dipende indirettamente da h_t attraverso la generazione del token intermedio — produce un termine di moto browniano:

$$dx_{\leq t} = (\hat{u} \circ \mathcal{f}(x_{\leq t}))\, dt + \sigma_t\, dW_t,$$

dove W_t è il moto browniano standard e σ_t è la deviazione standard degli errori accumulati al passo t. In virtù del Teorema di Donsker, quando $t \to \infty$ la somma $(1/\sqrt{t})\ \Sigma_{s \leq t}\ \varepsilon_s$ converge in distribuzione a $\mathcal{N}(0, \Sigma)$, e la magnitudine della distorsione soddisfa

$$\|h_t - h^*_t\|_2 \propto \sigma\sqrt{t},$$

dove $h^*_t = \mathcal{f}(x_{\leq t})$ è la traiettoria ottimale senza errori. Il cono di inferenza è la struttura geometrica che delimita probabilisticamente la traiettoria dei token generati: il suo raggio cresce come $\sigma\sqrt{t}$, e l'intersezione di questo cono con la traiettoria di un'altra sequenza — che causarebbe il collasso modale — ha probabilità crescente con σ. Ridurre σ tramite una loss ausiliaria riduce il raggio del cono e quindi la probabilità di collasso, il che costituisce la motivazione geometrica per la Semantic Tube Prediction.

III.11 — L'ipotesi geodetica e la tube semantica: definizioni formali e derivazioni

L'ipotesi geodetica afferma che le traiettorie error-free dei token seguono geodetiche su una varietà liscia nello spazio delle rappresentazioni latenti, e che queste geodetiche sono localmente lineari quasi ovunque. Più precisamente, una traiettoria temporale h^* si dice localmente lineare se esistono τ e ε tali che per qualsiasi tripla di indici $s < r < t$ con $|t - s| \le \tau$ si ha

$$\|(h_r - h_s)_{\perp}(h_t - h_s)\|_2 \le \varepsilon,$$

dove $x_{\perp y}$ denota la componente di x perpendicolare a y. Se la traiettoria fosse perfettamente lineare, questa componente sarebbe identicamente nulla per ogni r nel segmento $[s, t]$.

La nozione di segnale e rumore si cristallizza in questo contesto geometrico con precisione notevole. Dato il punto r intermedio tra s e t, il vettore $h_r - h_s$ si decompone in due parti:

$$(h_r - h_s) = (h_r - h_s)_{\parallel}(h_t - h_s) + (h_r - h_s)_{\perp}(h_t - h_s).$$

La componente parallela alla direzione geodetica è il segnale — la parte dell'incremento allineata con il percorso semantico inteso — e la componente perpendicolare è il rumore, ovvero la deviazione dalla traiettoria geodetica dovuta all'imperfetta convergenza del modello.

Il Lemma di Straightening (Lemma 3.2 di Huang et al.) fornisce il legame quantitativo tra la loss STP e la riduzione del rumore. Il lemma stabilisce: se $h_s = h_s$ e $h_t = h_t$ e se $\mathcal{L}_{STP} \le \varepsilon$ per ogni r con $s < r < t$, allora

$$\|(h_r - h_s)_{\perp}(h_t - h_s)\|_2 \le \sqrt{(2\varepsilon)} \cdot \|h_r - h_s\|_2.$$

La derivazione procede per via geometrica. Sia θ l'angolo tra $h_r - h_s$ e $h_t - h_s$, e sia θ' l'angolo tra $h_r - h_s$ e $h_t - h_r$. Si ha

$$\|(h_r - h_s)_{\perp}(h_t - h_s)\|_2 = \|h_r - h_s\|_2 \cdot \sin\theta.$$

Poiché $\theta' \ge \theta$ (il punto r è intermedio tra s e t), si ha $\sin\theta \le \sin\theta'$. La loss STP è

$$\mathcal{L}_{STP} = 1 - \cos(h_t - h_r, h_r - h_s),$$

donde $\mathcal{L}_{STP} = 1 - \cos\theta' \le \varepsilon$. Per ε sufficientemente piccolo $\cos\theta' \approx 1 - \theta'^2/2$, da cui $\theta'^2 \lesssim 2\varepsilon$ e quindi $\theta' \lesssim \sqrt{(2\varepsilon)}$. Per θ' piccolo $\sin\theta' \approx \theta'$, e si ottiene

$$\|(h_r - h_s)_{\perp}(h_t - h_s)\|_2 \le \|h_r - h_s\|_2 \cdot \sin\theta' \lesssim \sqrt{(2\varepsilon)} \cdot \|h_r - h_s\|_2. \quad \square$$

Il Teorema della Semantic Tube (Teorema 3.3) estende questo risultato locale a un enunciato globale: se h^* è localmente lineare e $\mathcal{L}_{STP} \to 0$ per ogni r con $0 \le s < r < t \le \tau$, allora

$$\|h_r - h^*\|_2 \lesssim \varepsilon,$$

dove $\|h - h^*\|_2 = \min_{r'} \|h_r - h_{r'}\|_2$ *è la distanza minima dal punto* h_r *alla traiettoria ottimale. La dimostrazione sfrutta il fatto che le condizioni al contorno* $h_0 = h_0$ e $h_\tau = h^*_\tau$ sono soddisfatte introducendo due token ausiliari ‹before-bos› e ‹after-eos›, applicando la disuguaglianza triangolare combinata con il Lemma di Straightening. La forma random della tube (Corollario 3.4) afferma che, quando $s < r < t$ sono scelti casualmente, si ha per qualsiasi ε e ε'

$$P(\|h_r - h^*\|_2 > \varepsilon + \varepsilon') \to 0 \text{ quando } \mathcal{L}_{STP} \to 0,$$

garantendo il confinamento nella tube con alta probabilità.

Il Lemma di Data Efficiency (Lemma H.1) fornisce il collegamento formale tra miglioramento del SNR e riduzione del numero di dati necessari. Modellando il processo come un canale gaussiano $X = Z + N$ con Z il segnale latente e $N \sim \mathcal{N}(0, \sigma^2 I)$, definendo $SNR = \mathbb{E}[\|Z\|^2] / \mathbb{E}[\|N\|^2]$ e usando la formula di capacità di Shannon $I(X; Y) = (1/2)\log(1 + SNR)$, il numero minimo di campioni m per raggiungere $H(Y \mid X^m) \leq \varepsilon$ soddisfa

$$m \geq (H(Y) - \varepsilon) / ((1/2)\log(1 + SNR)).$$

Questa è inversamente proporzionale a $\log(1 + SNR)$: aumentare il SNR riduce il numero di campioni necessari. Il complementare Lemma sull'accuratezza (Corollario H.3) stabilisce

$$P_e \gtrsim (H(Y) - m \cdot (1/2)\log(1 + SNR)) / \log |V|,$$

dove P_e è la probabilità di errore e $|V|$ è la dimensione del vocabolario. Se STP aumenta il SNR, P_e decresce a parità di m (migliore accuratezza) oppure il medesimo P_e si ottiene con m più piccolo (migliore data efficiency) — esattamente quanto si osserva sperimentalmente.

III.12 — Le Scaling Laws e la loro violazione strutturale

La conseguenza più significativa dei risultati empirici di STP per la teoria dell'apprendimento profondo è concettuale: l'ottenimento di prestazioni equivalenti con 1/16 dei dati viola il termine dati delle Scaling Laws in stile Chinchilla, che predicono una degradazione proporzionale a $D^{-\alpha_D}$ con $\alpha_D \approx 0.095$. Con 1/16 dei dati e stessa architettura, le leggi di Chinchilla prevedono una degradazione di $(16)^{0.095} \approx 1.35\times$ nella loss — ovvero una degradazione osservabile e non trascurabile nell'accuratezza — mentre STP mantiene l'accuratezza quasi inalterata.

La spiegazione è che le Scaling Laws caratterizzano l'addestramento tipico, non quello ottimale nel senso di massimo utilizzo dell'informazione per campione. L'obiettivo $\mathcal{L}_{NTP}$ standard ha un rapporto segnale-rumore strutturalmente limitato perché non distingue tra le variazioni semanticamente significative degli stati nascosti (componente parallela alla geodetica, segnale) e le variazioni ortogonali (rumore di training). STP migliora questo SNR senza aumentare il numero di parametri né il numero di token, il che significa che il bottleneck delle

Scaling Laws non è la quantità di compute per parametro o per token ma la qualità dell'informazione estratta da ciascun token. Le Scaling Laws misurano l'efficienza di uno specifico obiettivo di training ($\mathcal{L}$_NTP), non il limite fisico di quanta informazione semantica può essere estratta da un corpus di testo.

Questa osservazione riconnette con la distinzione fondamentale introdotta nel Capitolo I tra addestramento tipico e addestramento ottimale. I vincoli di compute, parametri e dati delle Scaling Laws sono vincoli empirici dell'addestramento condotto con $\mathcal{L}$_NTP su architetture Transformer standard. Modificando l'obiettivo — aggiungendo un termine geometrico che vincola la struttura delle traiettorie latenti — è possibile superare questi vincoli non aumentando la scala ma migliorando la qualità dell'apprendimento per ogni campione.

La connessione tra STP e l'architettura JEPA non è accessoria ma profonda. In I-JEPA il vincolo è sulla struttura predittiva locale (la rappresentazione di una regione deve essere predittibile da quella di un'altra regione nello stesso input); in STP il vincolo è sulla struttura dinamica globale (la traiettoria degli stati nascosti deve essere localmente retta lungo la sequenza). I-JEPA opera in un regime puramente spaziale; STP opera in un regime temporale-sequenziale. I-JEPA richiede un'architettura separata con encoder e predittore distinti; STP si riduce a un termine di loss computazionalmente quasi gratuito. Le due architetture sono dunque complementari, e la loro possibile integrazione — un LLM addestrato con I-JEPA per la struttura multi-view e con STP per la regolarizzazione delle traiettorie — rappresenta una direzione di sviluppo teoricamente motivata che potrebbe portare a guadagni cumulativi in efficienza dei dati superiori a quelli di ciascuna tecnica presa isolatamente.

III.13 — Connessioni tra JEPA, STP e i principi di invarianza nella fisica

La Semantic Tube Prediction introduce nella famiglia JEPA un principio che ha un profondo analogo nella fisica teorica: il principio di minima azione. In meccanica classica, la traiettoria fisicamente realizzata da un sistema è quella che rende stazionaria l'azione $S = \int L(q,\dot{q},t)dt$, dove L è la Lagrangiana. Le traiettorie fisicamente possibili non sono quelle che minimizzano l'energia istantanea, ma quelle che hanno la struttura geometrica ottimale globalmente — rette sul manifold della configurazione dotato della metrica appropriata.

La Geodesica Hypothesis di STP afferma qualcosa di strutturalmente analogo per le traiettorie degli stati nascosti negli LLM: le traiettorie error-free a convergenza sono geodetiche sul manifold semantico latente. La perdita $\mathcal{L}$_STP = 1 − cos(h_t − h_r, h_r − h_s) è una penalizzazione della curvatura locale della traiettoria — esattamente il tipo di penalizzazione che in fisica corrisponde alla minimizzazione dell'azione per una particella libera (Lagrangiana cinematica $L = (1/2)\|\dot{h}\|^2$). Il training di un LLM con $\mathcal{L}$_STP è quindi analogo alla selezione, tra tutte le traiettorie compatibili con i dati, di quelle con la struttura lagrangiana più semplice — le rette dello spazio semantico.

Questa connessione non è soltanto suggestiva: ha conseguenze pratiche. In fisica, i sistemi che minimizzano l'azione sono quelli più prevedibili e più stazionari rispetto a perturbazioni — caratteristica fondamentale per la stabilità. In apprendimento automatico, la regolarizzazione geodetica produce traiettorie latenti più stazionari rispetto alle variazioni del prompt di input, il che corrisponde a una maggiore robustezza della generazione rispetto a riformulazioni equivalenti. Il collegamento formale tra stabilità geodetica e robustezza alla riformulazione è quantificato dal Lemma di Straightening di STP: se $\mathcal{L}_STP \leq \varepsilon$, la deviazione dalla traiettoria ottimale è limitata da $\sqrt{(2\varepsilon)} \cdot \|h_r - h_s\|_2$.

Capitolo IV — World Models: Storia, Formalismo e Varianti Architetturali

IV.1 — L'idea di un modello interno del mondo: radici storiche

L'idea che un sistema intelligente debba possedere un modello interno del mondo — una rappresentazione della struttura causale dell'ambiente in cui opera, capace di simulare le conseguenze di azioni ipotetiche senza eseguirle fisicamente — ha radici che si estendono ben prima dell'era dell'apprendimento profondo. Kenneth Craik, psicologo scozzese, propose già nel 1943 nel suo "The Nature of Explanation" che il sistema nervoso costruisce "piccoli modelli" della realtà che usa per anticipare eventi e pianificare risposte. Richard Sutton, in un articolo del 1991 titolato "Dyna, an integrated architecture for learning, planning, and reacting", formalizzò questa intuizione nel contesto del reinforcement learning: un agente Dyna impara simultaneamente un modello del mondo (transizioni e reward) e una policy, e usa il modello per generare esperienze simulate che integrano il training basato su esperienza reale.

Ma è con il lavoro di Jürgen Schmidhuber — e in particolare con la serie di articoli sulla "Raumzeit-Kognition" (cognizione spaziotemporale) e sul "World Model" degli anni 1990 — che il concetto diventa formalmente preciso nel contesto delle reti neurali. Schmidhuber propose architetture dove una rete "controller" apprende a prevedere le conseguenze delle proprie azioni attraverso una rete "world model" appresa in modo autosupervisionato, e questo sistema integrato impara comportamenti complessi senza supervisione esterna. Queste idee anticipano di decenni molti degli sviluppi recenti, sebbene all'epoca il potere computazionale disponibile non permettesse di validarle su problemi di scala significativa.

Il contributo che ha definitivamente consacrato l'espressione "World Model" come categoria riconosciuta nella comunità moderna dell'apprendimento automatico è il paper di David Ha e Jürgen Schmidhuber del 2018, semplicemente intitolato "World Models". Questo lavoro, pur nella sua relativa semplicità architettonica rispetto agli sviluppi successivi, contiene in nuce le idee fondamentali che hanno guidato un'intera linea di ricerca fino ai modelli odierni.

Il filo storico che connette Craik (1943), Schmidhuber (anni '90), Ha e Schmidhuber (2018) e Hafner (2019-2023) è più coerente di quanto suggerisca la discontinuità cronologica. In tutti questi lavori, l'intuizione fondamentale è la stessa: un agente intelligente non reagisce al mondo come lo trova ma simula internamente le conseguenze delle proprie azioni prima di eseguirle, e aggiorna il suo modello interno sulla base del divario tra la simulazione e la realtà. Questa intuizione — che coincide con la struttura della percezione come inferenza inconscia di Helmholtz e con il principio di minima azione della fisica classica — riceve nel contesto delle reti neurali profonde la sua implementazione computazionale più sofisticata nella famiglia dei modelli DreamerV1-V3 di Hafner.

Ciò che distingue i World Models moderni dai loro predecessori non è l'intuizione di fondo — rimasta sostanzialmente invariata — ma la qualità delle rappresentazioni su cui il modello del mondo opera. Un World Model che opera su osservazioni sensoriali grezze deve modellare la varianza dell'intera distribuzione percettiva; un World Model che opera sullo spazio latente di un encoder JEPA può assumere che le varianti sensoriali irrilevanti siano già state soppresse, e concentrare la sua capacità rappresentativa sulla struttura causale effettiva del dominio. Questa è la connessione fondamentale tra JEPA e World Models che la struttura AMI rende esplicita: non si tratta di due famiglie di architetture parallele che possono coesistere separatamente, ma di due moduli funzionalmente complementari dove il primo prepara il terreno per il secondo.

IV.2 — Il modello di Ha e Schmidhuber: architettura e apprendimento

L'architettura del World Model del 2018 è strutturata in tre componenti distinti che corrispondono grosso modo alle funzioni di percezione, memoria e decisione:

Il **Variational Autoencoder (VAE)** V funge da encoder percettivo. Data un'osservazione sensoriale $x_t \in \mathbb{R}^{(H\times W\times 3)}$ (un frame), V produce un codice latente $z_t \in \mathbb{R}^k$ tramite un processo di inferenza variazionale. L'encoder del VAE $q_\varphi(z \mid x)$ produce i parametri di una distribuzione gaussiana diagonale:

$$q_\varphi(z \mid x) = \mathcal{N}(z; \mu_\varphi(x), \mathrm{diag}(\sigma^2_\varphi(x)))$$

dove $\mu_\varphi(x) \in \mathbb{R}^k$ e $\sigma^2_\varphi(x) \in \mathbb{R}^{k_+}$ sono le uscite della rete encoder con parametri φ. Un campione $z \sim q_\varphi(z \mid x)$ viene ottenuto tramite il reparametrization trick:

$$z = \mu_\varphi(x) + \sigma_\varphi(x) \odot \varepsilon, \varepsilon \sim \mathcal{N}(0, I)$$

dove $\odot$ denota il prodotto elemento per elemento (Hadamard). Il decoder del VAE $p_\theta(x \mid z)$ ricostruisce l'osservazione a partire dal codice latente, tipicamente come distribuzione gaussiana o Bernoulli a seconda del dominio.

La perdita di addestramento del VAE è l'Evidence Lower BOund (ELBO), che si deriva come segue. Si vuole massimizzare la log-verosimiglianza marginale $\log p_\theta(x) = \log \int p_\theta(x \mid z)\, p(z)\, dz$, dove $p(z) = \mathcal{N}(0, I)$ è il prior. Poiché questo

integrale è intrattabile, si usa la disuguaglianza di Jensen per ottenere un lower bound:

$$\log p_\theta(x) = \log \int p_\theta(x \mid z)\, p(z)\, dz$$

$$= \log \int [p_\theta(x \mid z)\, p(z) / q_\varphi(z \mid x)] \cdot q_\varphi(z \mid x)\, dz$$

$$\geq \int \log[p_\theta(x \mid z)\, p(z) / q_\varphi(z \mid x)] \cdot q_\varphi(z \mid x)\, dz$$

$$= \mathbb{E}_{z \sim q_\varphi(\cdot \mid x)} [\log p_\theta(x \mid z)] - KL(q_\varphi(z \mid x) \,\|\, p(z))$$

La disuguaglianza diventa uguaglianza quando $q_\varphi(z \mid x) = p_\theta(z \mid x)$ (la posterior vera). La perdita del VAE è dunque:

$$\mathcal{L}_{VAE}(\theta, \varphi) = - \mathbb{E}_{z \sim q_\varphi(\cdot \mid x)} [\log p_\theta(x \mid z)] + KL(q_\varphi(z \mid x) \,\|\, p(z))$$

Il primo termine è la perdita di ricostruzione (log-verosimiglianza negativa del decoder), il secondo è la divergenza KL tra la posterior approssimata e il prior, che per distribuzioni gaussiane ha forma chiusa:

$$KL(\mathcal{N}(\mu, \sigma^2 I) \,\|\, \mathcal{N}(0, I)) = (1/2) \cdot \Sigma_i [\, \mu_i^2 + \sigma_i^2 - 1 - \log \sigma_i^2 \,]$$

Il componente centrale del World Model di Ha e Schmidhuber è la **MDN-RNN** (Mixture Density Network – Recurrent Neural Network) M. Questa rete ha il compito di modellare la distribuzione condizionale del codice latente futuro dato il codice latente presente e l'azione:

$$p(z_{t+1} \mid z_t, a_t, h_t) \approx \Sigma_{i=1}^{K} \pi_i(h_t) \cdot \mathcal{N}(z_{t+1}; \mu_i(h_t), \sigma_i^2(h_t) \cdot I)$$

dove h_t è lo stato nascosto dell'LSTM (Long Short-Term Memory) che compone la parte ricorrente della MDN-RNN, aggiornato come:

$$h_t = LSTM(h_{t-1}, [z_t, a_t])$$

e $\pi_j(h_t)$, $\mu_j(h_t)$, $\sigma_j(h_t)$ sono le probabilità di mixing, medie e deviazioni standard del j-esimo componente gaussiano della mixture, prodotti da un layer lineare applicato a h_t. La distribuzione a mixture è necessaria perché la transizione nel mondo reale può essere intrinsecamente stocastica o multimodale: dati z_t e a_t, il futuro z_{t+1} potrebbe essere uno tra più possibili esiti distinti.

Il **controller** C è una rete semplice — nella versione originale un singolo layer lineare — che mappia la concatenazione del codice latente z_t e dello stato nascosto h_t all'azione a_t:

$$a_t = W_c \cdot [z_t, h_t] + b_c$$

I parametri di C vengono ottimizzati attraverso un algoritmo evolutivo (CMA-ES, Covariance Matrix Adaptation Evolution Strategy) piuttosto che backpropagation, il che permette di ottimizzare rispetto a reward cumulativi a lungo termine senza propagare gradienti attraverso la MDN-RNN.

La parte più visionaria del lavoro di Ha e Schmidhuber è la dimostrazione che il controller può essere addestrato interamente nella "mente" del modello — ovvero usando la MDN-RNN per generare roll-out simulati senza interazione con l'ambiente reale — e che il controller così addestrato può trasferirsi con successo all'ambiente reale. Questo è il primo esempio convincente di "imparare nel sogno" (learning in the dream), un'idea che Schmidhuber aveva proposto teoricamente decenni prima.

IV.3 — Il Recurrent State Space Model: formalismo completo

I limiti architetturali del modello di Ha e Schmidhuber — in particolare la separazione tra la codifica percettiva (VAE) e la modellazione dinamica (MDN-RNN) che impedisce l'apprendimento end-to-end di rappresentazioni ottimali per la predizione temporale — vengono superati nel framework DreamerV1 (Hafner et al., 2019) attraverso l'introduzione del Recurrent State Space Model (RSSM).

L'RSSM è un modello generativo latente che modella esplicitamente sia la struttura temporale che l'incertezza di stato, e che è progettato per essere addestrato end-to-end su sequenze di osservazioni e azioni. La sua struttura distingue due componenti dello stato latente: uno **stato deterministico** $h_t \in \mathbb{R}^{d_h}$ aggiornato ricorrentemente e uno **stato stocastico** $z_t \in \mathbb{R}^{d_z}$ campionato da una distribuzione condizionale.

Le equazioni del modello generativo RSSM sono le seguenti:

Equazione di transizione deterministica (prior della dinamica):

$$h_t = f_\theta(h_{t-1}, z_{t-1}, a_{t-1})$$

dove f_θ è tipicamente una GRU (Gated Recurrent Unit) con parametri θ. La GRU aggiorna lo stato nascosto secondo:

$$r_t = \sigma(W_r \cdot [h_{t-1}, [z_{t-1}, a_{t-1}]] + b_r)$$

$$u_t = \sigma(W_u \cdot [h_{t-1}, [z_{t-1}, a_{t-1}]] + b_u)$$

$$c_t = \tanh(W_c \cdot [r_t \odot h_{t-1}, [z_{t-1}, a_{t-1}]] + b_c)$$

$$h_t = (1 - u_t) \odot h_{t-1} + u_t \odot c_t$$

dove r_t è il reset gate, u_t è l'update gate, c_t è il candidate hidden state, e W_r, W_u, W_c, b_r, b_u, b_c sono i parametri della GRU.

Distribuzione prior dello stato stocastico:

$$p_\theta(z_t \mid h_t) = \mathcal{N}(z_t; \mu_\theta(h_t), \sigma^2_\theta(h_t) \cdot I)$$

dove $\mu_\theta(h_t)$ e $\sigma^2_\theta(h_t)$ sono prodotti da un MLP applicato allo stato deterministico h_t.

Distribuzione posterior dello stato stocastico (data l'osservazione):

$q_\varphi(z_t \mid h_t, x_t) = \mathcal{N}(z_t; \mu_\varphi(h_t, x_t), \sigma^2_\varphi(h_t, x_t) \cdot I)$

dove $\mu_\varphi(h_t, x_t)$ e $\sigma^2_\varphi(h_t, x_t)$ sono prodotti da un MLP che combina lo stato deterministico e una codifica dell'osservazione x_t.

Decoder delle osservazioni:

$p_\theta(x_t \mid h_t, z_t) = \mathcal{N}(x_t; \mu_{dec}(h_t, z_t), \sigma^2_{dec} \cdot I)$

oppure, per immagini, una distribuzione Laplaciana o un decoder convoluzionale che produce una distribuzione sui pixel.

Decoder del reward:

$p_\theta(r_t \mid h_t, z_t) = \mathcal{N}(r_t; \mu_r(h_t, z_t), \sigma^2_r \cdot I)$

Il modello viene addestrato massimizzando l'ELBO su sequenze di lunghezza T:

$$\text{ELBO} = \mathbb{E}_{q_\varphi}\left[\sum_{=1}^{T} \left(\log p_\theta(x_t \mid h_t, z_t) + \log p_\theta(r_t \mid h_t, z_t) - \beta \cdot \mathrm{KL}(q_\varphi(z_t \mid h_t, x_t) \,\|\, p_\theta(z_t \mid h_t)) \right) \right]$$

dove le aspettative sono calcolate rispetto alla distribuzione posterior q_φ (usando il reparametrization trick per il calcolo dei gradienti) e β è un coefficiente che bilancia la fedeltà della ricostruzione e la regolarizzazione del prior. La derivazione di questo ELBO dalla log-verosimiglianza marginale segue gli stessi passaggi della derivazione per il VAE, estesi al caso sequenziale attraverso l'applicazione ripetuta della disuguaglianza di Jensen lungo la catena temporale.

La proprietà cruciale dell'RSSM è che la componente deterministica h_t preserva le informazioni su tutto il passato in modo compresso ma senza perdita per le strutture rilevanti — grazie alla capacità della GRU di mantenere informazioni a lungo termine attraverso i suoi gate — mentre la componente stocastica z_t cattura l'incertezza residua che non può essere predetta deterministicamente dal passato e dall'azione. Questa separazione permette al modello di gestire correttamente ambienti parzialmente osservabili e stocastici, che sono la norma piuttosto che l'eccezione nel mondo reale.

IV.4 — DreamerV2: categorical latents e KL balancing

Le modifiche introdotte da DreamerV2 riflettono una comprensione più profonda delle difficoltà dell'ottimizzazione variazionale in sistemi sequenziali complessi. Prima di analizzare le soluzioni specifiche adottate, è utile comprendere le radici teoriche dei problemi che esse affrontano.

Il problema del posterior collapse nell'apprendimento variazionale sequenziale è un fenomeno ben noto nella letteratura dei Variational Autoencoders ma si manifesta in modo particolarmente acuto nei World Models perché coinvolge sequenze temporali dove l'informazione contestuale si accumula. Nel VAE

standard, il posterior collapse avviene quando l'ottimizzatore trova più conveniente ignorare le osservazioni (rendendo la posterior uguale al prior) che utilizzarle per aggiornare lo stato latente. Nei World Models sequenziali, il problema è amplificato dalla struttura ricorrente: se al passo t il sistema ha già collassato il posterior verso il prior, all'osservazione t+1 la componente deterministica h_t è già priva dell'informazione sul t-esimo step, e il posterior per t+1 si trova a dover fare il lavoro di recuperare informazione perduta — un compito impossibile in generale, che spinge verso un ulteriore collasso.

La soluzione del KL balancing affronta questo problema attraverso un principio elegante: separare il gradiente che fluisce verso il prior da quello che fluisce verso il posterior. Il prior viene "inseguito" dalla posterior con forza α, mentre il posterior "insegue" il prior con forza (1-α). Con $\alpha > 0.5$, si dà priorità all'aggiornamento del prior rispetto all'aggiornamento del posterior: il prior diventa un bersaglio mobile che si sposta verso le osservazioni correnti, invece di essere una distribuzione fissa che la posterior deve approssimare. Questo inverte la direzione del posterior collapse: invece di spingere la posterior verso un prior fisso, si spinge il prior verso la posterior informativa.

La scelta del valore $\alpha = 0.8$ in DreamerV2 è empiricamente motivata ma ha un'interpretazione teorica interessante. Con $\alpha = 0.8$, il gradiente che aggiorna il prior è quattro volte più forte di quello che aggiorna la posterior. Questo rapporto è scelto per bilanciare due requisiti conflittuali: il prior deve essere abbastanza flessibile da seguire le osservazioni senza resistenza eccessiva, ma abbastanza stabile da non oscillare caoticamente a ogni step. Il valore 0.8 rappresenta un punto di equilibrio empiricamente trovato che funziona bene su una varietà di ambienti, ma la ricerca di principi teorici che giustifichino questo valore specifico rimane aperta.

DreamerV2 (Hafner et al., 2020) introduce due modifiche architetturali fondamentali rispetto a DreamerV1

DreamerV2 (Hafner et al., 2020) introduce due modifiche architetturali fondamentali rispetto a DreamerV1: l'uso di variabili latenti categoriche anziché gaussiane, e una strategia di bilanciamento del termine KL nell'ELBO. Entrambe le modifiche sono motivate da analisi teoriche e si traducono in miglioramenti empirici sostanziali.

Riguardo alle variabili latenti categoriche: anziché modellare z_t come vettore gaussiano continuo, DreamerV2 lo modella come un insieme di L variabili categoriche indipendenti, ciascuna con K classi:

$z_t = (z_t^{(1)}, z_t^{(2)}, ..., z_t^{(L)}), z_t^{(l)} \in \{1, ..., K\}$ per ogni l

La distribuzione prior e posterior di $z_t^{(l)}$ sono distribuzioni categoriche (Categorical) i cui logit vengono prodotti da MLP applicati rispettivamente a h_t e a (h_t, enc(x_t)). In DreamerV2, $L = 32$ e $K = 32$, per un totale di $32 \times 32 = 1024$ bit effettivi di informazione stocastica per step temporale.

La motivazione per le variabili categoriche è duplice. Dal punto di vista computazionale, il campionamento da distribuzioni categoriche è più efficiente sul GPU rispetto alle gaussiane continue, e la distribuzione categorica ha un prior uniforme naturale che evita la necessità di calibrare la varianza del prior. Dal punto di vista teorico, la quantizzazione implicita nelle distribuzioni categoriche può migliorare la disentangling delle rappresentazioni: variabili che si "accendono" e "spengono" in modo discreto sono spesso più interpretabili di componenti gaussiane continue.

Il campionamento differenziabile da distribuzioni categoriche richiede il Gumbel-Softmax trick (o Straight-Through estimator). Il Gumbel-Softmax produce un campione "soft" differenziabile:

$\hat{z}_k = \exp((\log \pi_k + g_k) / \tau) / \Sigma \exp((\log \pi_j + g_j) / \tau)$

dove $g_k = -\log(-\log u_k)$ con $u_k \sim \text{Uniform}(0,1)$ sono variabili di Gumbel, π_k sono le probabilità della categorica e τ è la temperatura. Per $\tau \to 0$, $\hat{z}$ converge al campione one-hot discreto; per $\tau > 0$, $\hat{z}$ è una distribuzione morbida su cui backpropagation funziona normalmente. In DreamerV2 si usa $\tau = 1$ durante il training.

La seconda modifica di DreamerV2 è il KL balancing, che affronta il problema della "posterior collapse": quando il termine KL nell'ELBO è ottimizzato troppo aggressivamente (o quando β è troppo grande), il modello impara a ignorare le osservazioni e a usare solo il prior per generare le rappresentazioni, il che porta al collasso della posterior verso il prior. KL balancing introduce una ponderazione asimmetrica:

$KL(q_\varphi(z \mid h, x) \| p_\theta(z \mid h)) = \alpha \cdot KL(q_\varphi(z \mid h, x) \| sg(p_\theta(z \mid h))) + (1 - \alpha) \cdot KL(sg(q_\varphi(z \mid h, x)) \| p_\theta(z \mid h))$

con $\alpha = 0.8$. Il primo termine ottimizza la posterior per avvicinarsi al prior (con il prior fissato), il secondo ottimizza il prior per avvicinarsi alla posterior (con la posterior fissata). Pesare il primo termine più del secondo ($\alpha > 0.5$) significa dare priorità all'aggiornamento della posterior rispetto al prior, che empiricamente previene il posterior collapse mantenendo la posterior informativa rispetto alle osservazioni.

IV.5 — DreamerV3: la ricerca dell'algoritmo universale

La ricerca della "universalità algoritmica" — la capacità di un singolo algoritmo di funzionare bene su una varietà di domini senza tuning specifico — è uno degli obiettivi più ambiziosi e meno raggiunti del reinforcement learning. La storia del campo è costellata di algoritmi che producono risultati straordinari in contesti specifici e crolli spettacolari quando applicati a contesti leggermente diversi. DQN su Atari, PPO su MuJoCo, AlphaGo su Go: ciascuno di questi sistemi è ottimale nel suo dominio e mediocre o inutilizzabile fuori da esso.

Le ragioni di questa fragilità sono ben comprese a livello tecnico: diversi ambienti hanno scale di ricompensa diverse, orizzonti temporali diversi, densità di ricompensa diversa (alcune ricompense arrivano raramente, altre ad ogni step), dimensioni dello spazio delle azioni diverse. Ogni combinazione di queste caratteristiche richiede un'ingegneria specifica degli iperparametri — learning rate, gamma, entropy coefficient, clipping parameter — che non si trasferisce tra ambienti. La ricerca della "unica parametrizzazione" è in parte una ricerca di normalizzazioni che rendano questi parametri invarianti rispetto alle proprietà dell'ambiente.

DreamerV3 affronta sistematicamente ciascuna di queste fonti di variabilità. La symlog transformation normalizza le scale dei reward non attraverso un fattore di scala globale — che sarebbe semplice ma potrebbe comprimere eccessivamente la struttura locale dei reward — ma attraverso una funzione che comprime i valori grandi mantenendo la fedeltà ai valori piccoli. L'intuizione è che i reward molto grandi (come i punteggi negli Atari che possono raggiungere decine di migliaia) sono informativi sulla loro scala relativa ma non sulla loro scala assoluta: un reward di 10000 non è necessariamente "migliore" di un reward di 1000 in modo proporzionale alla loro differenza. La symlog trasforma questa differenza da lineare a logaritmica per i valori grandi, preservando la fedeltà per i valori piccoli dove la struttura locale è più informativa.

DreamerV3 (Hafner et al., 2023) compie un passo ulteriore verso la generalità

DreamerV3 (Hafner et al., 2023) compie un passo ulteriore verso la generalità, con l'obiettivo esplicito di costruire un algoritmo che funzioni su un'ampia gamma di ambienti, ricompense e scale senza richiedere tuning degli iperparametri specifico per dominio. Questo obiettivo — spesso chiamato "algorithm universality" nella letteratura di RL — è raro nella pratica perché la maggior parte degli algoritmi di RL richiede estensiva calibrazione per funzionare bene su nuovi ambienti.

Le innovazioni chiave di DreamerV3 sono tre: la symlog transformation, i free bits per il KL, e il percentile-based reward normalization.

La **symlog transformation** è definita come:

$$\text{symlog}(x) = \text{sign}(x) \cdot \log(|x| + 1)$$

e viene applicata sia alle osservazioni che ai reward e ai target del value function. La motivazione è che le reti neurali tendono a avere difficoltà nell'apprendere da valori con scale molto diverse: un reward che varia tra -10^4 e 10^4 richiede pesi molto diversi da un reward che varia tra -1 e 1. Applicando symlog si comprime uniformemente la scala dei valori grandi mantenendo il segno, il che rende l'apprendimento più stabile su ambienti con reward di scala arbitraria. In modo simmetrico, si usa:

$$\text{symexp}(x) = \text{sign}(x) \cdot (\exp(|x|) - 1)$$

come funzione inversa quando si vuole de-normalizzare le predizioni.

I **free bits** (o free nats) per il KL sono una tecnica per prevenire il posterior collapse in modo diverso dal KL balancing: anziché pesare i termini della KL, si vincola il minimo valore che la KL può assumere. Formalmente:

$$KL_clamped = \max(KL(q_\varphi(z \mid h, x) \parallel p_\theta(z \mid h)), F)$$

dove F è un valore minimo (in DreamerV3, F = 1 free nat per distribuzione categorica). Questo impedisce all'ottimizzatore di "spendere" capacità a ridurre la KL al di sotto di F, preservando la capacità informativa minima della posterior.

La **normalizzazione percentile del reward** trasforma il reward raw r_t in un reward normalizzato $\tilde{r}_t$ usando i percentili 5° e 95° dei reward visti finora:

$$\tilde{r}_t = r_t / \max(1, P_{95} - P_5)$$

dove P_5 e P_{95} sono i quantili stimati online. Questo rende l'algoritmo robusto a reward con scale molto diverse tra ambienti — da reward in centinaia di punti negli Atari game a reward in frazioni di metro negli ambienti di controllo continuo — senza richiedere normalizzazioni specifiche per dominio.

DreamerV3 con 200M parametri raggiunge prestazioni sovraumane su 44 dei 57 giochi Atari, il miglior score mai ottenuto da un singolo agente su PKU-SafeRL Bench, e prestazioni competitive su ambienti di controllo continuo come Humanoid (DMControl), tutto senza modifiche agli iperparametri. Questo risultato, pur essendo ancora lontano dall'AGI, rappresenta un passo qualitativo importante verso la generalità degli algoritmi di RL basati su World Models.

IV.6 — GAIA-1: World Models per la guida autonoma

L'applicazione dei World Models alla guida autonoma — uno dei problemi pratici più urgenti e più sfidanti dell'IA applicata — ha trovato nella startup britannica Wayve il suo punto di sintesi più sofisticato con il modello GAIA-1, presentato nel 2023.

GAIA-1 è un World Model autoregressivo di grande scala (9 miliardi di parametri) che impara a generare sequenze video realistiche di guida condizionate su azioni dell'ego-veicolo (accelerazione, sterzata, frenata), su condizioni meteo e sulle intenzioni di alto livello (svolta a destra, sorpasso, etc.). L'architettura è basata su un transformer autoregressivo che opera su token video discreti ottenuti tramite un tokenizer video (VQVAE), analogamente a come i Large Language Models operano su token testuali.

La novità concettuale di GAIA-1 rispetto ai World Models precedenti è duplice. In primo luogo, opera direttamente nello spazio delle osservazioni visive realistiche anziché in uno spazio latente compresso: le predizioni di GAIA-1 sono sequenze video perceptually realistic che possono essere visionate e valutate da ingegneri e tester umani. Questo è fondamentale per il contesto applicativo della guida autonoma, dove la comprensibilità e verificabilità delle predizioni del modello ha valore intrinseco per la sicurezza.

In secondo luogo, GAIA-1 è condizionabile su action tokens che non sono solo azioni di basso livello (angolo del volante, pressione del freno) ma anche segnali di alto livello (intenzione, tipo di manovra). Questo permette di usare GAIA-1 come "simulatore mentale" per pianificazione: il planner del sistema di guida autonoma può chiedere a GAIA-1 di simulare le conseguenze di diverse sequenze di azioni ipotetiche, valutarne la sicurezza e scegliere la più vantaggiosa — esattamente il tipo di pianificazione basata su World Model previsto dall'architettura di LeCun.

La perdita di addestramento di GAIA-1 è la cross-entropy autoregressiva standard sui token video:

$$\mathcal{L}_GAIA(\theta) = - \mathbb{E} \left[\Sigma_{=1}^{T} \log p_\theta(v_t \mid v_\{<t\}, a_\{<t\}, c) \right]$$

dove v_t è il token video al passo t, a_{<t} sono le azioni passate e c è il contesto condizionante (condizioni meteo, intenzione). Questa perdita è identica nella forma a quella di un LLM, il che permette di sfruttare tutta l'infrastruttura software e le tecniche di training scaling sviluppate per i modelli linguistici — un vantaggio pratico non trascurabile.

IV.7 — GameNGen e Genie: World Models per ambienti interattivi

Una linea parallela e convergente di sviluppo dei World Models riguarda la simulazione di videogiochi e ambienti interattivi. Il modello GameNGen, presentato da Google DeepMind nel 2024, è un diffusion model che impara a simulare il gioco DOOM a partire da una singola GPU, producendo frame successivi indistinguibili da quelli del motore di gioco originale con latenza inferiore a 50ms.

GameNGen è interessante non per la sua applicazione ludica ma per ciò che dimostra strutturalmente: un modello neurale può apprendere le regole fisiche e geometriche di un ambiente complesso direttamente da dati di traiettorie, senza mai aver avuto accesso al codice sorgente del simulatore. Le "leggi della fisica" di DOOM — la geometria degli spazi, le regole di collisione, la balistica delle armi, il comportamento dei nemici — emergono come proprietà apprese del modello di diffusione a partire dall'osservazione di migliaia di ore di gameplay.

Genie (Google DeepMind, 2024) spinge questo approccio un passo più avanti: è un World Model per ambienti interattivi 2D che non richiede annotazioni di azioni durante il training. Invece di apprendere la transizione s_t, a_t → s_{t+1} con azioni annotate, Genie impara a inferire latent action codes z_a_t dalla sequenza di osservazioni (s_t, s_{t+1}) attraverso un tokenizer di azioni non supervisionato, e poi usa questi codici latenti per condizionare la generazione di frame futuri. In altri termini, Genie impara spontaneamente la struttura delle azioni di controllo guardando come il mondo cambia tra frame consecutivi, senza che nessuno abbia detto al sistema cosa sono "su", "giù", "salta".

La perdita combinata di Genie è:

$\mathcal{L}$_Genie(θ_enc, θ_dec, θ_dyn) = $\mathcal{L}$_frame(θ_dec) + α · $\mathcal{L}$_action(θ_enc) + β · $\mathcal{L}$_commit(θ_enc, θ_dyn)

dove $\mathcal{L}$_frame è la perdita di predizione del frame successivo, $\mathcal{L}$_action è una perdita di regolarizzazione sull'azione latente (analoga al termine di commitment loss del VQVAE), e $\mathcal{L}$_commit è una perdita che allinea i codici di azione con le transizioni dinamiche.

IV.8 — Formalismo unificato e analisi comparativa

È utile a questo punto proporre un formalismo unificato che permetta di confrontare sistematicamente le diverse architetture di World Models analizzate. Tutti questi modelli possono essere scritti come istanze di un Partially Observable Markov Decision Process (POMDP) approssimato da una rete neurale.

Sia ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{O}$, T, O, R) un POMDP dove $\mathcal{S}$ è lo spazio degli stati latenti, $\mathcal{A}$ è lo spazio delle azioni, $\mathcal{O}$ è lo spazio delle osservazioni, T: $\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ è la funzione di transizione, O: $\mathcal{S} \to \Delta(\mathcal{O})$ è la funzione di osservazione e R: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ è la funzione di reward. Un World Model neurale impara un'approssimazione parametrica (T_θ, O_θ, R_θ) dei componenti (T, O, R) a partire da traiettorie osservate (o_1, a_1, r_1, o_2, a_2, r_2, ..., o_T).

Il modello di Ha e Schmidhuber corrisponde a un POMDP con spazio degli stati latenti $\mathcal{S}$ = ℝ^k × ℝ^d_h (codice VAE × stato LSTM), funzione di transizione modellata dalla MDN-RNN, e funzione di osservazione modellata dal VAE decoder. Il training avviene in due fasi separate: prima il VAE, poi la MDN-RNN condizionata sulle codifiche del VAE fissato.

L'RSSM di DreamerV1 corrisponde a un POMDP con $\mathcal{S}$ = ℝ^d_h × ℝ^d_z (stato deterministico × stato stocastico), con la particolarità che la componente deterministica h_t è aggiornata deterministicamente dalla GRU mentre la componente stocastica z_t è campionata da una distribuzione condizionata a h_t. Il training è end-to-end attraverso la massimizzazione dell'ELBO sequenziale.

DreamerV2 e DreamerV3 estendono questo schema con variabili latenti categoriche e tecniche di stabilizzazione del training. GameNGen e GAIA-1 usano fondamentalmente la funzione di transizione come modello di generazione visiva, operando in spazi di token discreti piuttosto che in spazi latenti continui.

La differenza fondamentale tra queste varianti non è strutturale ma riguarda il tipo di inductive bias incorporato: l'RSSM incorpora l'inductive bias che lo stato del mondo ha componenti deterministiche (prevedibili dal passato) e stocastiche (intrinsecamente imprevedibili); i modelli autoregressivi su token discreti incorporano l'inductive bias che il mondo può essere descritto sequenzialmente attraverso un vocabolario finito di "atomi percettivi". Entrambe le scelte hanno pregi e limiti che si manifestano in diversi tipi di ambienti e compiti.

Un aspetto cruciale che distingue i World Models di successo è la capacità di gestire efficacemente il "compounding error" nella generazione di traiettorie lunghe: piccoli errori nella predizione di z_{t+1} si accumulano lungo la sequenza, portando a traiettorie che deviano sempre più dalla dinamica reale. L'RSSM mitiga questo problema attraverso la componente stocastica che mantiene la varianza delle predizioni calibrata rispetto all'incertezza reale; i modelli autoregressivi su video lo affrontano attraverso tecniche di augmentation temporale durante il training (history dropout, random clipping) che rendono il modello robusto rispetto a condizioni iniziali imperfette. In entrambi i casi, il problema del compounding error rimane una sfida aperta, particolarmente critica per applicazioni di pianificazione su orizzonti temporali lunghi come la guida autonoma.

Il capitolo successivo si occuperà di come JEPA e i World Models si integrano nella visione sintetica di LeCun dell'Autonomous Machine Intelligence, analizzando sia le connessioni architetturali che le difficoltà che rimangono da superare prima che tale visione possa essere realizzata.

IV.9 — La struttura ODE/SDE delle traiettorie latenti e il fondamento teorico della decomposizione deterministica-stocastica dell'RSSM

La trattazione delle architetture World Model nei paragrafi IV.2–IV.5 aveva messo in luce la distinzione fondamentale dell'RSSM tra componente deterministica h_t — aggiornata ricorrentemente dalla GRU secondo h_t = f_θ(h_{t-1}, z_{t-1}, a_{t-1}) — e componente stocastica z_t — campionata dalla distribuzione condizionale p_θ(z_t | h_t) nel regime prior o q_φ(z_t | h_t, x_t) nel regime posterior. Questa decomposizione era stata presentata come una scelta ingegneristica motivata da considerazioni pratiche: la componente deterministica preserva le informazioni sul passato in modo compresso ma senza perdita per le strutture rilevanti, mentre quella stocastica cattura l'incertezza residua imprevedibile. Il quadro era empiricamente giustificato ma teoricamente incompleto: non si era offerta una derivazione formale del perché le traiettorie dei sistemi dinamici apprese debbano naturalmente decomporsi in questo modo, né di quali siano le proprietà delle componenti risultanti. Il paper di Huang, LeCun e Balestriero (2026) sulla Semantic Tube Prediction colma questa lacuna in modo sorprendentemente diretto, derivando la stessa struttura ODE/SDE dal comportamento convergente di un modello autoregressivo sotto l'obiettivo di predizione del token successivo.

Il punto di partenza è la modellazione formale della dinamica di addestramento. Sia $x_{\leq t} \in \mathbb{R}^{T \times d_{model}}$ la sequenza di embedding dei token fino al tempo t, e sia la dinamica di training descritta dall'equazione di differenza:

$$x_{\leq t+1} = x_{\leq t} \oplus \hat{u} \circ f(x_{\leq t}),$$

dove f è la funzione del network convergente, $\hat{u}$ è la funzione di unembedding, e $\oplus$ denota la concatenazione. Definita l'operazione di prefix-removal $\ominus$ tramite la

funzione di lift v(x, t) = [0, …, 0, x|{*pos t+1*}, *0, …, 0*] ∈ ℝ^{*T*×*d_model*}, *si ha* *x*{≤t+1} ⊖ x_{≤t} = û ◦ f(x_{≤t}), che è della forma z_{t+1} - z_t = g(z_t, t) — la struttura standard di un'ODE a tempo discreto. Nel limite di passo infinitesimo, si ottiene la Proposizione 2.1 di STP:

dx_{≤t} = û ◦ f(x_{≤t}) dt,

ovvero la Proposizione che descrive il Training ODE. Il termine û ◦ f(x_{≤t}) corrisponde esattamente alla componente deterministica dell'RSSM: dato lo stato corrente della sequenza (l'intero contesto x_{≤t}, che svolge il ruolo dell'hidden state deterministico h_t nella GRU), il passo successivo è completamente determinato dalla funzione del network convergente. La GRU nell'RSSM implementa questo update deterministico su un contesto compresso (h_t = f_θ(h_{t-1}, z_{t-1}, a_{t-1})), mentre nel formalismo STP l'update deterministico opera sull'intero prefisso x_{≤t} — una scelta computazionalmente più costosa ma teoricamente più precisa.

La componente stocastica emerge a tempo di inferenza. Durante l'addestramento con teacher forcing, l'errore residuo ε_t = h_t - h_*t (dove ht = f(x*{≤t}) è la traiettoria ottimale) non si propaga: il token successivo viene fornito dal ground truth indipendentemente dall'errore corrente. Durante l'inferenza, invece, h_{t+1} dipende da h_t tramite il meccanismo autoregressivo, e ε_t si accumula. Nell'approssimazione di rete a larghezza infinita (Yang e Littwin, 2021), le pre-attivazioni convergono a processi gaussiani, e ε_t è modellabile come rumore bianco gaussiano i.i.d. Il teorema di Donsker garantisce che la somma cumulativa di ε_s per s ≤ t converge, dopo normalizzazione appropriata, a un moto browniano. Questo porta alla Proposizione B.1 di STP:

dx_{≤t} = û ◦ f(x_{≤t}) dt + σ_t dW_t,

dove σ_t dW_t è il termine browniano con intensità σ_t. La connessione con la distribuzione prior dell'RSSM è immediata: p_θ(z_t | h_t) = 𝒩(z_t; μ_θ(h_t), σ²_θ(h_t) · I) modella esattamente questa componente stocastica — la distribuzione condizionale dello stato successivo dato lo stato deterministico corrente, tenendo conto dell'incertezza residua che la componente deterministica non può predire. La varianza σ²_θ(h_t) prodotta dall'MLP dell'RSSM è un'approssimazione appresa di σ_t² nell'SDE, e il fatto che DreamerV2 usi variabili latenti categoriche (L = 32, K = 32) anziché gaussiane continue può essere interpretato come una discretizzazione del processo stocastico continuo nell'SDE — una scelta di quantizzazione che migliora la stabilità computazionale a scapito della fedeltà matematica all'SDE continua.

La Proposizione F.1 di STP quantifica il comportamento del termine stocastico nel tempo: la deviazione ‖h_t - h_*t‖₂ *tra la traiettoria realizzata e quella ottimale cresce come* *σ√t, ovvero il "cono di inferenza" si allarga con la radice quadrata del numero di passi. Nell'RSSM, questo si manifesta come la crescente incertezza nelle predizioni del World Model per orizzonti temporali lunghi: con K step nel futuro, la varianza cumulativa della distribuzione predittiva scala approssimativamente come K · σ² (per la somma di K varianze i.i.d.) o come σ²K*

(per processi autocorrelati), il che corrisponde al compounding error delle traiettorie simulate in DreamerV3 discusso in IV.7 della monografia. La dipendenza dalla radice quadrata $\|h_t - h_t\|_2 \propto \sigma\sqrt{t}$ nell'SDE suggerisce che il compounding error nell'RSSM dovrebbe crescere più lentamente che lineare nel numero di passi — compatibile con le osservazioni empiriche che DreamerV3 mantiene prestazioni ragionevoli su orizzonti di 15 passi (Atari) ma degrada su orizzonti molto più lunghi.

La connessione teorica tra STP e RSSM non si esaurisce nella struttura ODE/SDE ma si estende alla questione di come ridurre σ_t, ovvero come migliorare la qualità delle predizioni del World Model per orizzonti lunghi. STP afferma che minimizzare $\mathcal{L}$_STP durante il training del Perception Module riduce l'ampiezza dell'errore residuo ε_t nell'LLM, abbassando σ_t nel termine stocastico dell'SDE. Se si applica lo stesso principio al World Model — aggiungendo un termine $\mathcal{L}$_STP calcolato sulle traiettorie degli stati latenti ŝ_t nel training dell'RSSM — si otterrebbe una regolarizzazione geometrica che forza le traiettorie latenti del World Model a essere localmente lineari, riducendo il compounding error per orizzonti lunghi senza richiedere modifiche architetturali al modello di transizione. Questa possibilità non è stata sperimentata alla data di stesura di questa monografia, ma è teoricamente ben motivata: la Geodesica Hypothesis di STP si applica a qualsiasi sequenza di rappresentazioni latenti dove il Principio di Minima Azione governa la traiettoria ottimale, e le traiettorie degli stati latenti in un World Model ben addestrato dovrebbero soddisfare questo principio nella misura in cui il World Model apprende la struttura causale del dominio.

IV.10 — Il cono di inferenza come metrica di incertezza e la sua relazione con l'incertezza epistemica di AMI

La formalizzazione del compounding error come cono di inferenza di raggio $\|h_t - h^*_t\|_2 \propto \sigma\sqrt{t}$ *fornisce una metrica concreta per quantificare l'incertezza del World Model che era rimasta implicita nella trattazione della sezione IV.7 e nella discussione dell'incertezza epistemica nella sezione V.5 del Capitolo V. Il cono di inferenza è la proiezione su un singolo punto della distribuzione stocastica* $dx_{\leq t} = \hat{u} \circ f(x_{\leq t})\, dt + \sigma_t\, dW_t$: esso descrive la regione dello spazio latente in cui la traiettoria generata può trovarsi con alta probabilità dopo t passi, data la condizione iniziale.

La densità di probabilità della posizione nel cono è approssimativamente gaussiana con varianza $\sigma^2 t$ nella direzione radiale (normale alla traiettoria ottimale), il che permette di calcolare la probabilità di mode collapse come la probabilità che il cono dell'agente A intersechi la geodetica dell'agente B:

$$P(\text{mode collapse}) \approx P(\|h_t^A - h_t^A\|_2 + \|h_t^B - h_t^B\|_2 > \|h_t^A - h_t^B\|_2) \approx P(\sigma_A \sqrt{t} + \sigma_B \sqrt{t} > \|h_t^A - h_t^B\|_2 / \sqrt{t} \cdot \sqrt{t}) = P((\sigma_A + \sigma_B) > \|h_t^A - h_t^B\|_2 / \sqrt{t})$$

Questa espressione mostra che il mode collapse è improbabile quando la distanza tra le geodetiche ottimali è grande rispetto al prodotto $(\sigma_A + \sigma_B)$, e diventa sempre più probabile all'aumentare di t per una data separazione iniziale. Per il

World Model di AMI, questa analisi si traduce in una stima dell'orizzonte massimo di pianificazione affidabile: se due traiettorie di pianificazione distinte hanno geodetica inizialmente separate di una distanza d_0, il loro cono di inferenza si sovrappone dopo circa t_collapse ≈ (d_0 / (2σ))² passi. Ridurre σ attraverso ℒ_STP allarga questo orizzonte di un fattore ϱ²_STP — per ϱ_STP ≈ 4 (radice di 16) in termine di σ, l'orizzonte di pianificazione affidabile si estende di un fattore 16. Questa è la connessione diretta tra l'efficienza dei dati di STP (riduzione di ϱ_STP ≈ 16 nel numero di campioni necessari) e la qualità del World Model (allargamento di un fattore ϱ²_STP nell'orizzonte di pianificazione affidabile): entrambi gli effetti emergono dalla stessa riduzione dell'ampiezza del rumore σ_t nelle traiettorie latenti.

Per l'architettura AMI complessiva, questa analisi modifica la stima dell'orizzonte di pianificazione pratico discussa nella sezione V.3 sul piano gerarchico. Se σ_t è ridotto di un fattore ϱ_σ rispetto al baseline (addestramento senza ℒ_STP), l'orizzonte temporale su cui il World Model produce predizioni affidabili si estende di ϱ²_σ — un guadagno quadratico nella scala di pianificazione a partire da un miglioramento lineare nel SNR. Questo suggerisce che l'incorporazione di ℒ_STP nel training del Perception Module e del World Model potrebbe avere impatti sproporzionati sulla capacità di pianificazione a lungo termine dell'agente AMI rispetto all'impatto diretto sull'efficienza dei dati, trasformando un miglioramento quantitativo (dati necessari) in un miglioramento qualitativo (profondità del ragionamento temporale).

Capitolo V — L'Architettura dell'Intelligenza Autonoma: JEPA, World Models e il Framework AMI

V.1 — Il manifesto di LeCun: una lettura tecnica

Nel maggio 2022 Yann LeCun pubblica su OpenReview un documento di settantadue pagine intitolato "A Path Towards Autonomous Machine Intelligence" che costituisce, al momento della sua stesura, l'articolazione più sistematica mai tentata da un singolo ricercatore di una visione architettonica completa per un sistema di intelligenza artificiale generale. Il documento non è un articolo empirico nel senso tradizionale — non riporta esperimenti con numeri comparabili a benchmark standard — ma è qualcosa di più raro e ambizioso: un programma di ricerca teoricamente motivato che identifica i principi fondamentali che un sistema AGI deve soddisfare e propone architetture concrete per realizzarli.

Leggere «A Path Towards Autonomous Machine Intelligence» dopo cinque anni dalla sua pubblicazione — con i successivi sviluppi di LeJEPA, della Semantic Tube Prediction e di V-JEPA 2 — produce un senso di straordinaria profeticità tecnica. Molte delle proposte di LeCun che nel 2022 sembravano visionarie o speculative hanno ricevuto implementazioni concrete nel periodo successivo, anche se non sempre nella forma esatta che LeCun aveva anticipato. Il documento

programmatico del 2022 andrebbe letto non come un manuale di istruzioni — alcune delle sue proposte concrete sono state superate dagli sviluppi successivi — ma come una mappa delle tensioni concettuali che il campo dell'intelligenza artificiale deve risolvere per progredire verso sistemi genuinamente generali. In questo senso, il suo valore non diminuisce con il tempo ma aumenta: identificare le domande giuste è spesso più difficile che risponderle, e LeCun ha avuto il merito di articolare con chiarezza inusuale l'insieme di domande che una teoria dell'intelligenza artificiale generale deve rispondere.

Una lettura superficiale potrebbe ridurlo a una collezione di intuizioni e proposte speculative. Una lettura attenta rivela invece una struttura argomentativa coerente che parte da critiche ben motivate ai paradigmi esistenti (apprendimento supervisionato, apprendimento per rinforzo model-free, modelli generativi) e costruisce da queste critiche un sistema alternativo i cui componenti sono ciascuno motivato dalla necessità di risolvere problemi specifici identificati nella sezione precedente. Il documento è, in questo senso, un esempio di architettura cognitiva nell'accezione più nobile del termine: non una lista di desiderata ma una struttura causale dove ciascun componente esiste perché senza di esso qualcosa di specifico non funziona.

Il framework che LeCun propone viene da lui denominato Autonomous Machine Intelligence (AMI) ed è strutturato attorno a sei moduli funzionali che interagiscono attraverso un'architettura message-passing su grafi energetici. Nei paragrafi che seguono si analizza ciascuno di questi moduli nella sua specificità tecnica, si mostrano le connessioni con le architetture JEPA e World Model analizzate nei capitoli precedenti, e si valuta la plausibilità delle proposte alla luce dello stato attuale della ricerca.

V.2 — I sei moduli dell'Autonomous Machine Intelligence

Il sistema AMI comprende sei componenti funzionali che LeCun identifica come necessari per un agente capace di apprendimento e comportamento autonomo nel mondo reale.

Il **Perception Module** (Modulo Percettivo) ha il compito di costruire una rappresentazione parziale dello stato del mondo s_t a partire dalle osservazioni sensoriali x_t. Questa rappresentazione non è e non dovrebbe essere una copia fedele dell'input sensoriale, ma un'astrazione che cattura le strutture rilevanti per la pianificazione e l'azione. I-JEPA è il candidato naturale per questo modulo: impara rappresentazioni latenti di alta semanticità attraverso la predizione latente, produce embedding che catturano relazioni strutturali piuttosto che dettagli sensoriali, e si addestra in modo autosupervisionato senza richiedere annotazioni umane. La connessione tra il Perception Module e I-JEPA è dunque diretta e non metaforica: I-JEPA è progettato come implementazione del Perception Module.

Il **World Model** (Modulo di Modellazione del Mondo) predice le conseguenze di azioni ipotetiche nello spazio delle rappresentazioni latenti. Formalmente, dato lo

stato latente corrente ŝ_t e un'azione ipotetica ã, produce una distribuzione sul prossimo stato latente:

$$p(\hat{s}_{t+1} \mid \hat{s}_t, \tilde{a}) \approx WM_\theta(\hat{s}_t, \tilde{a})$$

Questo modulo corrisponde direttamente all'RSSM di Dreamer o ai World Models discussi nel capitolo precedente. La distinzione cruciale rispetto al loro uso standard in DreamerV3 è che in AMI il World Model opera nello spazio delle rappresentazioni prodotte dal Perception Module (I-JEPA), non direttamente nello spazio delle osservazioni sensoriali. Questa separazione tra percezione e modellazione del mondo — suggerita in modo vago nel framework DreamerV1 attraverso il VAE ma non implementata come modulo distinto con apprendimento separato — è teoricamente vantaggiosa perché permette di ottimizzare ciascun modulo rispetto a obiettivi specifici: I-JEPA per la qualità delle rappresentazioni, il World Model per la fedeltà della predizione dinamica.

I sei moduli del framework AMI non sono elementi indipendenti che possono essere progettati separatamente e poi connessi: sono componenti di un sistema co-evolutivo dove la qualità di ciascun modulo dipende dalla qualità degli altri. Questa interdipendenza è uno dei principi più profondi del framework e anche una delle sue sfide implementative più difficili.

Consideriamo la dipendenza tra il Perception Module e il World Model. Il Perception Module produce le rappresentazioni ŝ_t che il World Model usa per fare predizioni. Se le rappresentazioni di I-JEPA sono buone — semanticamente ricche, geometricamente regolari, informativamente complete rispetto allo stato del mondo — il World Model deve apprendere una funzione di transizione relativamente semplice sullo spazio latente. Se invece le rappresentazioni sono povere — rumorose, ridondanti, o dipendenti da caratteristiche sensoriali irrilevanti — il World Model deve compensare questa povertà, apprendendo trasformazioni che separino l'informazione rilevante da quella irrilevante mentre si occupa simultaneamente di modellare la dinamica. Questa doppia responsabilità degrada inevitabilmente la qualità della modellazione dinamica.

La dipendenza nella direzione opposta è ugualmente importante. Se il World Model è accurato, il Perception Module può essere addestrato in un regime dove le sue predizioni vengono corrette da segnali informativi sul futuro. Se il World Model è impreciso, i segnali di correzione sono rumorosi e il Perception Module apprende rappresentazioni che si adattano al rumore del World Model piuttosto che alla struttura del mondo. Questo loop di co-dipendenza suggerisce che i due moduli dovrebbero idealmente essere addestrati in modo coordinato — non in sequenza — per evitare che le imperfezioni dell'uno vincoli le possibilità dell'altro fin dall'inizio.

Il **Cost Module** (Modulo di Costo) computa un costo scalare

Il **Cost Module** (Modulo di Costo) computa un costo scalare c(ŝ, ã) che misura quanto indesiderabile è uno stato-azione a partire dal punto di vista dell'agente. In

LeCun il costo non è un reward esterno fornito dall'ambiente (come nel RL tradizionale) ma la somma di due contributi: un **intrinsic cost** IC(ŝ) che codifica obiettivi biologicamente (o artificialmente) predefiniti come la sopravvivenza, il mantenimento dell'omeostasi, la soddisfazione dei bisogni primari; e un **trainable critic** V_φ(ŝ) appreso attraverso l'interazione con l'ambiente che stima il costo futuro scontato da uno stato.

Il **Short-Term Memory Module** (Modulo di Memoria a Breve Termine) mantiene uno stato interno che accumula informazioni rilevanti sull'ambiente nel corso di un episodio. Tecnicamente, questo modulo può essere implementato come un buffer di stati latenti recenti, come un Transformer che attende su una window di osservazioni passate, o come un memory-augmented network. La distinzione dal World Model è che la Short-Term Memory preserva informazioni sull'episodio corrente senza fare predizioni esplicite, mentre il World Model usa queste informazioni per simulare futuri possibili.

L'**Actor Module** (Modulo Attoriale) produce una distribuzione su azioni da eseguire. In AMI, contrariamente agli approcci RL standard dove il policy network mappa direttamente stati in azioni, l'Actor è subordinato al processo di pianificazione: le azioni vengono selezionate ottimizzando una sequenza ipotetica di azioni che minimizzi il costo cumulativo atteso secondo il World Model. Questo introduce un loop fondamentale: Actor → World Model → Cost Module → Actor.

Il **Configurator Module** (Modulo Configuratore) ha il compito di configurare gli altri moduli in modo appropriato al contesto e agli obiettivi di alto livello dell'agente. LeCun lo descrive come un sistema che seleziona quali parametri del World Model e del Cost Module siano rilevanti per il compito corrente, come un meccanismo di attenzione di ordine superiore che distribuisce le risorse computazionali in modo adattivo. Tecnicamente, questo modulo potrebbe essere implementato come un sistema di routing dinamico o come un insieme di "expert subnetworks" che vengono composte in modo compito-dipendente.

V.3 — Il Model Predictive Control come principio architetturale

L'innovazione più profonda del framework AMI rispetto agli approcci RL standard non è la lista dei moduli ma la proposta di usare il Model Predictive Control (MPC) come meccanismo di pianificazione, piuttosto che apprendere una policy end-to-end tramite backpropagation del reward.

La scelta del Model Predictive Control come meccanismo di pianificazione in AMI riflette una comprensione profonda delle differenze tra pianificazione e apprendimento della policy. Nella tradizione del reinforcement learning end-to-end, la policy $\pi: \mathcal{S} \to \Delta(\mathcal{A})$ è un oggetto appreso che mappa stati in distribuzioni di probabilità sulle azioni. Una volta appresa, la policy è un meccanismo reattivo: l'azione viene prodotta in tempo costante dalla valutazione di $\pi(s)$ senza alcuna simulazione del futuro. Questo è il suo punto di forza

computazionale e il suo punto di debolezza concettuale: la policy non "pensa" al futuro; riflette la struttura del futuro implicita nei parametri appresi durante il training.

Il MPC rovescia questa architettura: invece di una policy che riflette implicitamente il futuro, si ha un World Model che lo simula esplicitamente. A ogni passo temporale, l'agente esegue un ciclo di pianificazione che considera una finestra di H azioni future, simula le loro conseguenze attraverso il World Model, valuta le traiettorie risultanti attraverso il Cost Module, e seleziona l'azione iniziale della sequenza ottimale. Solo dopo aver eseguito questa simulazione l'agente agisce.

Il costo computazionale di questo approccio — H forward pass attraverso il World Model a ogni step di azione — è ovviamente superiore a quello di una policy reattiva. Ma questo costo è giustificato da benefici qualitativi che non possono essere ottenuti da policy reattive: la capacità di adattarsi online a nuove informazioni sull'ambiente, la possibilità di incorporare vincoli espliciti sulle azioni ammissibili, la trasparenza del ragionamento (ogni azione viene scelta con una traiettoria futura esplicita), e la robustezza rispetto a distributional shift moderato (se l'ambiente è cambiato rispetto al training, il World Model può essere aggiornato online e la nuova policy emerge automaticamente dal MPC senza riaddestramento).

Il MPC è una tecnica di controllo ottimale sviluppata originariamente per sistemi di processo industriale

Il MPC è una tecnica di controllo ottimale sviluppata originariamente per sistemi di processo industriale negli anni 1980, che trova ora una nuova vita come principio architetturale per i sistemi di IA. L'idea fondamentale è: anziché apprendere una policy $\pi: \mathcal{S} \to \Delta(\mathcal{A})$ che mappa stati in azioni in modo diretto, si risolve online un problema di ottimizzazione che trova la sequenza di azioni a_0, a_1, ..., a_H−1 che minimizza il costo cumulativo su un orizzonte H secondo il World Model:

a*_0, ..., a*_{H−1} = argmin_{a_0,...,a_{H−1}} Σ₌₀^{H−1} c(ŝ_t, a_t) + c_terminal(ŝ_H)

soggetto a: ŝ_{t+1} = WM_θ(ŝ_t, a_t), ŝ_0 = s_current

Solo la prima azione a*_0 viene eseguita, poi il problema viene re-ottimizzato al passo successivo con osservazione aggiornata — da cui il termine "receding horizon control". Il vantaggio del MPC rispetto all'apprendimento di policy è che: (1) permette adattamento online senza riaddestrare la policy; (2) esplicita la struttura del ragionamento (ogni azione viene scelta con previsione esplicita delle conseguenze); (3) permette di incorporare vincoli espliciti sulle azioni ammissibili in modo più naturale.

Il problema dell'ottimizzazione MPC in spazi di azioni continui può essere risolto con tecniche di ottimizzazione differenziabile — backpropagating attraverso il

World Model per calcolare $\partial c / \partial a_t$ — oppure con tecniche di campionamento come il Cross-Entropy Method (CEM) o il Model Predictive Path Integral (MPPI). La scelta tra approcci differenziabili e campionamento dipende dalla struttura della funzione di costo e dello spazio delle azioni.

LeCun enfatizza che il MPC basato su World Model permette di affrontare un problema che è fondamentale per l'AGI ma che non ha ancora ricevuto soluzione soddisfacente nei sistemi attuali: la pianificazione gerarchica su scale temporali multiple. Un agente che deve navigare da Milano a Tokyo deve pianificare simultaneamente a scale diverse — dalla scelta del volo intercontinentale alla gestione del passo di una porta — e queste scale non sono indipendenti: la scelta di alto livello determina il contesto in cui avvengono le scelte di basso livello.

Il MPC gerarchico in AMI sarebbe strutturato come una sequenza di moduli di pianificazione con orizzonti e risoluzioni temporali diverse. A livello più alto, il planner produce un piano astratto che specifica una sequenza di stati obiettivo latenti $(\tilde{s}_1, \tilde{s}_2, ..., \tilde{s}_K)$ su scala temporale lunga. A livello intermedio, il planner produce per ciascun segmento tra stati obiettivi consecutivi una sequenza di azioni di medio livello. A livello più basso, il controller trasforma le azioni di medio livello in segnali di controllo motorio di basso livello. Ciascun livello usa il proprio World Model calibrato sulla scala temporale appropriata.

V.4 — Grafi energetici e ottimizzazione per inferenza

Una caratteristica distintiva del framework AMI che lo separa dalla maggior parte degli approcci di apprendimento automatico è il suo fondamento sulle **Energy-Based Models (EBM)**: l'interazione tra i moduli non è modellata come passaggio di messaggi probabilistici espliciti (come in una Bayesian Network) ma come minimizzazione di una funzione di energia globale definita sull'insieme delle variabili del sistema.

Un Energy-Based Model definisce una funzione di energia $F_\theta: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ che assegna un'energia scalare a ogni configurazione di variabili di input (x) e di variabili latenti (y). L'inferenza consiste nel trovare le configurazioni y* che minimizzano $F_\theta(x, y^*)$ per un dato input x:

$$y^* = \text{argmin}_y F_\theta(x, y)$$

L'apprendimento dei parametri θ consiste nel fare sì che le configurazioni "corrette" (compatibili con la semantica del dominio) abbiano energia bassa e le configurazioni "scorrette" abbiano energia alta. La perdita standard per gli EBM è la **contrastive loss**:

$$\mathcal{L}_{EBM}(\theta) = \mathbb{E}_{(x, y^+)} [F_\theta(x, y^+)] - \mathbb{E}_{(x, y^-)} [F_\theta(x, y^-)]$$

dove y^+ sono le variabili latenti corrette (campionate dalla distribuzione reale) e y^- sono le variabili latenti scorrette (campionate da una distribuzione contrastiva). Varianti includono la hinge loss:

$\mathcal{L}$_EBM_hinge(θ) = $\mathbb{E}$ [max(0, m + F_θ(x, y^+) − F_θ(x, y^-))]

dove m è un margine minimo tra le energie dei campioni positivi e negativi.

Nel contesto AMI, la funzione di energia globale del sistema è definita come somma delle energie locali dei singoli moduli:

F_global(x, ŝ, a, ĉ) = F_perception(x, ŝ) + F_world(ŝ, a, ŝ') + F_cost(ŝ, a, ĉ) + F_memory(ŝ, m)

dove ŝ è la rappresentazione dello stato latente, a è l'azione, ĉ è il costo e m è il contenuto della memoria. L'inferenza consiste nel trovare simultaneamente le assegnazioni di tutti questi variabili latenti che minimizzano F_global dato l'input sensoriale x — un problema di ottimizzazione congiunta che in generale richiede tecniche iterative come la discesa del gradiente alternata o algoritmi di message-passing su grafi.

Questa prospettiva unifica elegantemente la percezione (inferenza dello stato latente che minimizza F_perception), la pianificazione (ricerca della sequenza di azioni che minimizza la somma di F_world e F_cost lungo un orizzonte), e l'apprendimento (aggiornamento dei parametri θ per abbassare le energie dei percorsi corretti e alzare quelle dei percorsi scorretti). La complessità computazionale di questa ottimizzazione congiunta è in generale elevata, il che motiva le approssimazioni architetturali discusse in precedenza (separazione dei moduli, MPC con orizzonte finito, approssimazione con Actor network).

V.5 — La questione della modellazione dell'incertezza

Un aspetto che il framework AMI affronta in modo più esplicito rispetto ai World Models standard è la modellazione dell'incertezza epistemica — l'incertezza derivante dalla mancanza di conoscenza, distinta dall'incertezza aleatorica che è intrinseca alla stocasticità del mondo.

In un agente che esplora un ambiente sconosciuto, l'incertezza epistemica dovrebbe guidare l'esplorazione: l'agente dovrebbe preferire azioni che riducono la propria incertezza sul modello del mondo (esplorazione) rispetto ad azioni che massimizzano il reward atteso secondo il modello corrente (sfruttamento). Questo trade-off esplorazione-sfruttamento è uno dei problemi fondamentali del reinforcement learning.

LeCun propone che il Cost Module incorpori esplicitamente un termine di **epistemic cost** che penalizza gli stati del mondo dove l'incertezza predittiva del World Model è elevata. Formalmente, se p_θ(ŝ_{t+1} | ŝ_t, a_t) è la distribuzione predittiva del World Model, l'incertezza epistemica può essere misurata dall'entropia di questa distribuzione:

H(p_θ(ŝ_{t+1} | ŝ_t, a_t)) = − ∫ p_θ(ŝ_{t+1} | ŝ_t, a_t) · log p_θ(ŝ_{t+1} | ŝ_t, a_t) dŝ_{t+1}

Un agente orientato all'esplorazione può aggiungere al costo intrinseco un termine:

$$IC_epistemic(\hat{s}_t, a_t) = -\lambda \cdot H(p_\theta(\hat{s}_{t+1} \mid \hat{s}_t, a_t))$$

che premia azioni che portano in regioni di alta incertezza del World Model, inducendo comportamenti di esplorazione attiva. Tuttavia, la misura dell'incertezza epistemica richiede un modello probabilistico ben calibrato del World Model, il che è tecnicamente difficile per reti neurali di grande scala. Metodi pratici includono l'ensemble di World Models (dove l'incertezza è stimata dalla varianza delle predizioni dei diversi componenti dell'ensemble) o le varianti Bayesian (dropout come approximate inference, normalizing flows per distribuzioni posterior sui pesi).

La questione dell'incertezza si intreccia con il problema del **distributional shift**: un agente addestrato in un certo ambiente e deployato in uno leggermente diverso si trova ad operare in regioni dello spazio degli stati dove il suo World Model è poco affidabile, il che può portare a comportamenti catastrofici. La gestione robusta del distributional shift richiede che l'agente sia consapevole di quando si trova "fuori distribuzione" — ovvero in regioni dello spazio degli stati dove le sue rappresentazioni e predizioni non sono affidabili — e che in questi casi si comporti in modo conservativo piuttosto che pianificare aggressivamente.

V.6 — Apprendimento per imitazione e il bootstrap dell'AMI

Una questione pratica fondamentale che il framework AMI deve affrontare è: come si inizializza un sistema così complesso? Un agente AMI a pieno regime ha bisogno di un World Model pre-addestrato, di rappresentazioni percettive pre-addestrate, di un Cost Module calibrato e di una Short-Term Memory operativa. Ma per addestrare il World Model serve un agente che esplori l'ambiente, il che richiede un controller funzionante, il che richiede un World Model... Il problema dell'inizializzazione è un chicken-and-egg che non ha soluzione banale.

LeCun propone di risolvere questo problema attraverso una combinazione di **apprendimento per imitazione** e di **curriculum learning**. L'apprendimento per imitazione — più specificamente l'Inverse Reward Design o il Behavioral Cloning su dimostrazioni di esperti umani — permette di inizializzare la policy (Actor) e il World Model in modo grossolano ma sufficiente per cominciare l'esplorazione autonoma. Il curriculum learning prevede di esporre l'agente inizialmente ad ambienti semplici dove l'esplorazione casuale produce esperienze sufficientemente ricche, e di aumentare progressivamente la complessità dell'ambiente man mano che l'agente migliora.

Formalmente, l'apprendimento per imitazione tramite Behavioral Cloning ottimizza:

$$\mathcal{L}_BC(\pi_\theta) = \mathbb{E}_{(s, a) \sim D_expert} [-\log \pi_\theta(a \mid s)]$$

dove D_expert è un dataset di traiettorie esperte. Questo approccio soffre del problema del **covariate shift**: la distribuzione degli stati visitati dalla policy appresa può divergere dalla distribuzione degli stati nelle dimostrazioni, il che porta a errori di compounding simili al compounding error dei World Models. L'algoritmo DAgger (Dataset Aggregation) mitiga questo problema raccogliendo iterativamente nuove dimostrazioni esperte negli stati visitati dalla policy corrente, ma richiede l'accesso continuo all'esperto.

Nel contesto di AMI, l'apprendimento per imitazione non viene usato per apprendere una policy finale ma per inizializzare il sistema con una conoscenza di base del mondo che poi viene raffinata attraverso l'apprendimento autosupervisionato (I-JEPA per le rappresentazioni, World Model per la dinamica) e l'apprendimento per rinforzo (Cost Module per gli obiettivi). Questa strategia di bootstrap è concettualmente elegante ma difficile da implementare in pratica a causa della sensibilità della fase di inizializzazione alla qualità delle dimostrazioni e agli iperparametri del curriculum.

V.7 — Connessioni con la Predictive Coding e il cervello

Il framework AMI non è costruito nel vuoto rispetto alle neuroscienze computazionali: LeCun fa riferimento esplicito alla tradizione della Predictive Coding come ispirazione per alcune scelte architetturali, in particolare per il meccanismo di feedback top-down che permette ai livelli superiori della gerarchia di modulare l'elaborazione nei livelli inferiori.

La teoria della Predictive Coding (PC), sviluppata sistematicamente da Karl Friston nel contesto del suo framework della Energia Libera (Free Energy), propone che il cervello sia fondamentalmente un sistema che minimizza la "sorpresa" — ovvero il log-likelihood negativo delle osservazioni sensoriali sotto il modello generativo interno — attraverso un processo iterativo di inferenza variazionale. Ogni livello della gerarchia corticale mantiene sia una stima dello stato corrente (rappresentazione) che una predizione di ciò che il livello inferiore dovrebbe osservare (predizione); i segnali di feedback trasportano le predizioni dall'alto verso il basso, i segnali di errore di predizione viaggiano dal basso verso l'alto, e l'apprendimento avviene aggiornando sia le rappresentazioni che i parametri del modello per minimizzare gli errori di predizione.

La connessione con AMI è strutturale: il World Model in AMI è il modello generativo che genera predizioni sugli stati latenti futuri, mentre le rappresentazioni di I-JEPA corrispondono alle attivazioni delle aree sensoriali nel framework PC. Il segnale di predizione fluisce dall'alto (World Model) verso il basso (Perception Module), e l'errore di predizione — la differenza tra lo stato latente predetto dal World Model e quello osservato dall'encoder — fluisce verso l'alto per aggiornare sia il World Model che l'encoder.

Tuttavia, la connessione non è perfetta. Il framework PC di Friston richiede un processo iterativo di minimizzazione della energia libera ad ogni step temporale

(non solo durante l'apprendimento), il che corrisponde computazionalmente a un'iterazione di inferenza variazionale ad ogni osservazione. I sistemi AMI e DreamerV3 non implementano questa inferenza iterativa online ma usano un encoder ammortizzato (inference network) che produce stime dello stato latente in un singolo forward pass — un'approssimazione efficiente ma che perde alcune proprietà teoriche del framework PC. La direzione di ricerca che integra inferenza iterativa con deep learning è attiva e promettente ma non ha ancora prodotto sistemi scalabili paragonabili a DreamerV3 in termini di prestazioni.

V.8 — Limitazioni del framework AMI e questioni aperte

Un'analisi critica onesta del framework AMI deve identificare le questioni che rimangono aperte o che il documento di LeCun affronta in modo incompleto, distinguendo tra quelle che sono difficoltà ingegneristiche (risolvibili con più risorse e ricerca) e quelle che sono problemi di principio (che richiedono nuove idee concettuali).

La prima questione è la **composizionalità sistematica**. Un sistema AMI con un World Model pre-addestrato può pianificare in ambienti simili a quelli di training, ma generalizzare a nuove combinazioni di oggetti, agenti e relazioni mai viste durante l'addestramento — la cosiddetta generalizzazione sistematica o composizionale — richiede che lo spazio latente abbia struttura algebrica esplicita che permetta la composizione. Le rappresentazioni di I-JEPA sono distribuite (ogni oggetto è codificato come un vettore denso) anziché simboliche (ogni oggetto ha un identificatore discreto separato), il che rende difficile garantire la composizionalità. La connessione con le reti neurali relazionali e i grafi neurali è qui rilevante ma non ancora integrata nel framework AMI in modo convincente.

La seconda questione è il **catastrophic forgetting** nell'apprendimento continuo. Un agente AMI che incontra nuovi ambienti o nuovi compiti deve aggiornare il proprio World Model e le proprie rappresentazioni senza dimenticare ciò che ha imparato in precedenza. Il catastrophic forgetting — il fenomeno per cui l'apprendimento di nuove informazioni degrada le prestazioni su informazioni apprese in precedenza — è ancora un problema aperto nelle reti neurali standard. Tecniche come l'Elastic Weight Consolidation (EWC) o i Progressive Neural Networks mitigano il problema ma non lo risolvono completamente.

La terza questione — forse la più fondamentale — è il **radicamento del significato** (symbol grounding). Le rappresentazioni di I-JEPA e gli stati latenti del World Model sono vettori di numeri che acquisiscono significato solo in relazione ai dati su cui il sistema è stato addestrato. Ma il significato, nel senso in cui è rilevante per la pianificazione e il ragionamento adattativi, richiede un collegamento tra le rappresentazioni interne e le loro conseguenze causali nel mondo reale — collegamento che si stabilisce solo attraverso l'azione e la percezione dell'agente embodied nel mondo. Questo problema — noto come symbol grounding problem nella filosofia della mente cognitiva — non trova

soluzione nel framework AMI nella sua forma attuale, dove il sistema è addestrato prevalentemente su dati visivi passivi senza embodiment attivo.

V.8.A — Il problema della scala temporale: pianificazione gerarchica e World Models a granularità multipla

Una delle sfide più profonde del framework AMI è la pianificazione su scale temporali radicalmente diverse. Un agente che deve preparare una cena complessa deve simultaneamente pianificare l'ordine delle portate (scala di ore), la sequenza dei piatti per ciascuna portata (scala di decine di minuti), i passaggi di ciascuna ricetta (scala di minuti), e i movimenti motori fini per tagliare, mescolare e cucinare (scala di secondi e microsecondi). Nessun singolo World Model con una risoluzione temporale fissa può gestire questa gerarchia di scale in modo efficiente.

Il framework AMI risponde a questa sfida attraverso la pianificazione gerarchica, ma la sua realizzazione concreta richiede una gerarchia di World Models calibrati su scale temporali diverse. Al livello più alto, un World Model "strategico" opera su rappresentazioni compresse di stati del mondo a lungo termine — "stare in cucina", "stare servendo la cena" — con transizioni che coprono decine di minuti. Al livello intermedio, un World Model "tattico" opera su sequenze di azioni di medio livello — "tagliare le verdure", "accendere il fuoco" — con transizioni che coprono minuti. Al livello più basso, un World Model "motorio" opera su traiettorie di controllo articolare — "estendere il braccio", "stringere la presa" — con transizioni che coprono decimi di secondo.

Il nesso tra questi livelli è problematico sia tecnicamente che teoricamente. Sul piano tecnico, i World Models a diversi livelli devono condividere informazione sullo stato del mondo in modo coerente: le azioni del World Model strategico devono vincolare il contesto in cui il World Model tattico opera, e viceversa i fallimenti del World Model tattico devono essere comunicati al livello strategico per aggiornare la pianificazione di lungo periodo. Sul piano teorico, non è chiaro quale sia il numero "giusto" di livelli gerarchici, quale granularità temporale sia appropriata per ciascuno, e come apprendere automaticamente la struttura gerarchica ottimale piuttosto che imporla manualmente.

V.8.B — Il problema dell'orizzonte e la decadenza dell'informazione nel planning

Un aspetto tecnico del MPC che il framework AMI affronta ma che merita approfondimento è il deterioramento della qualità delle predizioni con l'aumento dell'orizzonte di pianificazione H. Questo deterioramento — noto come "compounding error" — è causato dall'accumulo degli errori di predizione del World Model lungo la sequenza: se l'errore al passo t è ε_t, al passo t+1 il World Model opera su uno stato errato $\hat{s}_t + \varepsilon_t$, producendo un errore ε_{t+1} che dipende sia dall'errore intrinseco del World Model che dall'errore di input. In generale, gli errori si accumulano più che linearmente, e per orizzonti sufficientemente lunghi le predizioni diventano inaffidabili.

La formalizzazione del compounding error richiede ipotesi sulla struttura della funzione di transizione del World Model. Se WM_θ: $\mathcal{Z} \times \mathcal{A} \to \mathcal{Z}$ ha costante di Lipschitz L rispetto allo stato (con $\|WM_\theta(z_1, a) - WM_\theta(z_2, a)\| \leq L \cdot \|z_1 - z_2\|$), allora l'errore dopo H step cresce come L^H · ε₀, dove ε₀ è l'errore iniziale. Per L < 1 (contrattivo), l'errore decresce — il World Model è stabile. Per L > 1 (espansivo), l'errore cresce esponenzialmente — il World Model è instabile. Per la maggior parte dei World Models pratici, L ≈ 1 con fluttuazioni intorno all'unità a seconda della regione dello spazio degli stati, producendo un compounding error che cresce approssimativamente come √H per orizzonti moderati.

Il framework AMI mitiga questo problema attraverso la pianificazione gerarchica: invece di pianificare su un orizzonte H con il World Model completo, si pianifica su un orizzonte H/K con il World Model ad alto livello (che opera su rappresentazioni compresse con dinamiche più lente e quindi errori di compounding inferiori) e poi su K orizzonti di lunghezza 1 con il World Model a basso livello. Questo schema divide il compounding error da L^H a K · L^{H/K}, una riduzione esponenziale per L > 1.

V.9 — LeJEPA e il fondamento variazionale degli Energy-Based Models: dall'euristico al provato

La trattazione degli Energy-Based Models nel paragrafo V.4 aveva messo in luce la struttura energetica del framework AMI con la sua funzione globale

F_global(x, ŝ, a, ĉ) = F_perception(x, ŝ) + F_world(ŝ, a, ŝ') + F_cost(ŝ, a, ĉ) + F_memory(ŝ, m),

sottolineando come l'inferenza corrisponda alla minimizzazione congiunta di questa quantità rispetto alle variabili latenti, e l'apprendimento all'aggiornamento dei parametri θ per abbassare le energie delle configurazioni corrette. Ciò che mancava alla trattazione era una risposta precisa a una domanda la cui risposta sembrava inevitabilmente circolare: quale distribuzione dovrebbero avere le rappresentazioni prodotte dal Perception Module — implementato tramite I-JEPA — affinché la minimizzazione dell'energia globale sia computazionalmente ben condizionata e statisticamente ottimale per compiti downstream arbitrari? La ricerca precedente aveva eluso questa domanda, sostituendo la risposta con euristiche: stop-gradient per prevenire il collasso, reti teacher-student con schedule EMA per la stabilità, bilanciamento manuale di iperparametri. LeJEPA (Balestriero e LeCun, 2025) trasforma la domanda da retorica a formale e la risponde in modo rigoroso, con conseguenze profonde per il progetto AMI nel suo insieme.

Il punto di partenza è una riformulazione del problema dell'apprendimento autosupervisionato all'interno del linguaggio degli EBM. Un encoder $f_\theta : \mathbb{R}^D \to \mathbb{R}^K$ produce, per ogni osservazione x, una codifica latente $z = f_\theta(x)$. Si assume che z sia la variabile latente di un EBM implicito in cui la funzione di energia è $F(x, z) = \|z - f_\theta(x)\|^2 / 2$, il che corrisponde a un prior gaussiano isotrópico sulle rappresentazioni. L'inferenza produrrebbe banalmente $z^* = f_\theta(x)$ — ma questa

osservazione, lungi dall'essere triviale, rivela che la distribuzione indotta sull'embedding da F deve soddisfare condizioni di ottimalità rispetto a compiti downstream. La domanda diventa: quale forma funzionale di p_Z minimizza il rischio su una famiglia ampia di tali compiti, sotto un vincolo scalare sulla covarianza?

La risposta — che LeJEPA deriva rigorosamente per probing lineare (sezione 3.1) e non lineare (sezione 3.2) — è la distribuzione gaussiana isotrópica $\mathcal{N}(0, (\varkappa/K) \cdot I_K)$, dove $\varkappa = \mathrm{tr}(\mathrm{Cov}(Z))$. La dimostrazione per il probing lineare si articola in due lemmi complementari.

Il primo lemma riguarda il bias dell'estimatore. Si consideri l'estimatore ai minimi quadrati con regolarizzazione di Tikhonov:

$$\hat{\beta} = (Z^\top Z + \lambda_{wd} I)^{-1} Z^\top y,$$

con $Z \in \mathbb{R}^{N \times K}$ la matrice degli embedding, $y \in \mathbb{R}^N$ il vettore target e $\lambda_{wd} \geq 0$ il coefficiente di regolarizzazione. La decomposizione spettrale $G = Z^\top Z / N = Q \Lambda Q^\top$ con $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ porta a:

$$\mathrm{Bias}(\hat{\beta}) = \mathbb{E}[\hat{\beta}] - \beta^* = -\lambda_{wd} (G + \lambda_{wd} I)^{-1} \beta^* = -\lambda_{wd} Q (\Lambda + \lambda_{wd} I)^{-1} Q^\top \beta^*.$$

Per una configurazione anisotropica in cui i λ_k non sono tutti uguali, si scelga $\beta^* = \varkappa_0 \cdot q_K$, dove q_K è l'autovettore relativo al più piccolo autovalore λ_K. Le norme del bias per la configurazione anisotropica e per quella isotrópica, a parità di energia totale $\Sigma_k \lambda_k = N\varkappa$, sono rispettivamente:

$$\|\mathrm{Bias}(\hat{\beta})\|_{\mathrm{aniso}} = \lambda_{wd} / (\lambda_K + \lambda_{wd}) \cdot \|\beta^*\|, \quad \|\mathrm{Bias}(\hat{\beta})\|_{\mathrm{iso}} = \lambda_{wd} / (\bar{\varkappa} + \lambda_{wd}) \cdot \|\beta^*\|,$$

con $\bar{\varkappa} = \Sigma_k \lambda_k / K = N\varkappa/K$ la varianza media per dimensione. Poiché $\lambda_K < \bar{\varkappa}$ ogniqualvolta la distribuzione sia strettamente anisotropica (almeno due autovalori distinti), si ha

$$\lambda_{wd} / (\lambda_K + \lambda_{wd}) > \lambda_{wd} / (\bar{\varkappa} + \lambda_{wd}),$$

stabilendo che la configurazione anisotropica amplifica il bias per quella specifica scelta di β^*. Questo non è un risultato nel caso peggiore per un compito avverso: è una garanzia che per ogni distribuzione non isotrópica esiste sempre almeno un compito downstream per cui l'estimatore è più distorto rispetto al caso isotrópico. L'intuizione geometrica è nitida: le dimensioni latenti con varianza ridotta ricevono una penalizzazione sproporzionata dalla regolarizzazione, che non distingue tra direzioni con diversa copertura statistica.

Il secondo lemma riguarda la varianza. Con $\lambda_{wd} = 0$ l'estimatore OLS è non distorto, e la sua covarianza è:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (Z^\top Z)^{-1} = \sigma^2 Q \Lambda^{-1} Q^\top,$$

con varianza totale $\mathrm{tr}(\mathrm{Var}(\beta)) = \sigma^2 \sum_{k=1}^{K} 1/\lambda_k$. La convessità stretta di $x \mapsto 1/x$ su $(0, \infty)$ e la disuguaglianza di Jensen forniscono immediatamente:

$$\frac{1}{K}\sum_k \frac{1}{\lambda_k} \geq \frac{1}{\frac{1}{K}\sum_k \lambda_k} = \frac{K}{N\varkappa/K} = \frac{K^2}{N\varkappa},$$

con uguaglianza se e solo se tutti i λ_k coincidono. La varianza totale è dunque minimizzata dalla distribuzione isotrópica:

$$\mathrm{tr}(\mathrm{Var}(\beta))_{\mathrm{aniso}} \geq \mathrm{tr}(\mathrm{Var}(\beta))_{\mathrm{iso}} = \frac{\sigma^2 K^2}{N\varkappa}.$$

I due lemmi congiuntamente stabiliscono l'ottimalità della distribuzione isotrópica rispetto al bias-variance tradeoff per qualsiasi compito downstream lineare.

Il passo successivo dell'argomentazione di LeJEPA è l'identificazione della forma funzionale precisa della distribuzione: non basta sapere che la covarianza deve essere scalata dall'identità, occorre specificare quale famiglia di distribuzioni marginali p_Z minimizza il rischio downstream. A tal fine si usa il funzionale di informazione di Fisher:

$$J(p) = \int_{\mathbb{R}^K} \|\nabla \log p(z)\|^2 \, p(z) \, dz,$$

che misura la curvatura media del log-likelihood e governa la precisione con cui i parametri di posizione possono essere stimati. Mediante il bound di Cramér-Rao nella forma matriciale — che afferma $\mathcal{I}(\theta) \succeq \Sigma^{-1}$ con $\mathcal{I}(\theta)$ la matrice di Fisher della famiglia di location — si dimostra $J(p) \geq \mathrm{tr}(\Sigma^{-1})$, con uguaglianza se e solo se il score $\nabla \log p(z)$ è lineare in z, ovvero $p = \mathcal{N}(0, \Sigma)$. Minimizzare $\mathrm{tr}(\Sigma^{-1})$ sotto i vincoli scalari sulla covarianza (traccia fissa, determinante fisso, norma di Frobenius fissa, raggio spettrale limitato) porta in tutti i casi a $\Sigma = s\, I_K$ per la costante scalare s appropriata, la cui formula esplicita dipende dal tipo di vincolo:

— sotto vincolo di traccia $\mathrm{tr}(\Sigma) = t$: minimo a $\Sigma = (t/K)\, I_K$, con $J_{\min} = K^2/t$; — sotto vincolo di determinante $\det(\Sigma) = \delta$: minimo a $\Sigma = \delta^{1/K}\, I_K$, con $J_{\min} = K\, \delta^{-1/K}$; — sotto vincolo di norma di Frobenius $\|\Sigma\|_F = c$: minimo a $\Sigma = (c/\sqrt{K})\, I_K$, con $J_{\min} = K^{3/2}/c$; — sotto vincolo di raggio spettrale $\varrho(\Sigma) \leq r$: minimo a $\Sigma = r\, I_K$, con $J_{\min} = K/r$.

In tutti i casi la distribuzione ottimale è la gaussiana isotrópica con la costante scalare determinata dal vincolo. Questo risultato è provato per il probing lineare tramite i due lemmi sopra, e viene esteso al probing non lineare — sia con k-NN radiale sia con regressione kernel di Nadaraya-Watson — attraverso una derivazione dello squared integrated bias (ISB). Per il k-NN, il bias puntuale all'interrogazione q è:

$$\mathrm{Bias}(q) = r_0^2/(K+2) \cdot [\nabla \eta(q)^T \nabla \log p(z)(q) + (1/2)\Delta\eta(q)] + o(r_0^2),$$

dove $\eta : \mathbb{R}^K \to \mathbb{R}$ è la funzione target, r_0 è il raggio del vicinato, e $\nabla \log p(z)$ è lo score della densità degli embedding. L'ISB integrato sulla distribuzione di query $q \sim p_Z$ è:

$\mathbb{E}_q[\text{Bias}(q)^2] = (r_0^2/(K+2))^2 \cdot \tau^2_g\, J(p) + O(r_0^4)$,

dove $\tau^2_g = \text{tr}(\text{Cov}_\eta(\nabla\eta(z)\nabla\eta(z)^\top)) / K$ misura l'isotropia media del gradiente della funzione target. Minimizzare l'ISB rispetto a p significa minimizzare J(p), e il minimizzatore è ancora la gaussiana isotrópica, il che estende l'ottimalità al regime non lineare.

La rilevanza di questi risultati per il framework AMI va oltre la semplice validazione teorica dell'encoder. Il Perception Module in AMI ha il compito di produrre rappresentazioni $\hat{s}_t$ che alimentano tutti gli altri moduli — in particolare il World Model, il Cost Module e lo Short-Term Memory. Se la distribuzione di $\hat{s}_t$ è anisotropica, il World Model opera in uno spazio dove alcune direzioni latenti hanno varianza molto più elevata di altre: il modello di transizione $WM_\theta(\hat{s}_t, a)$ sarà mal condizionato lungo le direzioni a bassa varianza, producendo predizioni imprecise proprio nelle regioni dello spazio latente meno rappresentate nella distribuzione dei dati. Un Perception Module che produce embedding gaussiani isotrópici non è una preferenza estetica ma una necessità funzionale per la buona condizionalità del sistema integrato.

La connessione tra LeJEPA e gli EBM è quindi bidirezionale: LeJEPA deriva la struttura ottimale degli embedding dall'analisi del rischio downstream, mentre il framework EBM di AMI richiede esattamente quella struttura per il corretto funzionamento dell'inferenza congiunta su F_global. L'implementazione di questa struttura ottimale è affidata a SIGReg (Sketched Isotropic Gaussian Regularization), che introduce un nuovo tipo di vincolo energetico nel training: anziché penalizzare direttamente la distanza dalla distribuzione target in alta dimensione — operazione computazionalmente costosa con complessità quadratica nel numero di campioni — SIGReg proietta gli embedding lungo M direzioni casuali $a \in \mathcal{S}^{K-1}$ e confronta le distribuzioni univariate proiettate con quelle della gaussiana standard tramite il test di Epps-Pulley:

$\text{SIGReg}_T(\mathbb{A}, \{f_\theta(x_n)\}\{n=1\}^N) \triangleq (1/|\mathbb{A}|) \Sigma\{a\in\mathbb{A}\} T(\{a^\top f_\theta(x_n)\}_\{n=1\}^N)$,

dove T è la statistica di Epps-Pulley:

$EP = N \int_\{-\infty\}^\{\infty\} |\hat{\varphi}_X(t) - \varphi(t)|^2\, w(t)\, dt$,

con $\hat{\varphi}_X(t) = (1/n)\, \Sigma_\{j=1\}^n e^\{itX_j\}$ la funzione caratteristica empirica e $\varphi(t) = e^\{-t^2/2\}$ la funzione caratteristica della gaussiana standard, e $w(t) = e^\{-t^2/\sigma^2\}$ la finestra gaussiana con $\sigma = 1$.

La scelta del test di Epps-Pulley rispetto a test basati sui momenti (Jarque-Bera esteso) o sulla funzione di distribuzione cumulativa (Cramér-von Mises, Anderson-Darling) è motivata da ragioni di stabilità e scalabilità che il Teorema 4 di LeJEPA formalizza. Per i test basati sui momenti, il gradiente rispetto al k-esimo momento scala come O(k) e la varianza delle stime Monte Carlo cresce come $O(k^2 m_\{2(k-1)\})$, il che rende instabile l'addestramento per $k \geq 3$. Per i test basati sulla

CDF, la necessità di ordinare campioni rompe il parallelismo intrinseco della SGD su multi-GPU. Il test di Epps-Pulley, invece, gode di gradienti uniformemente limitati: per ogni campione z_i,

$$|\partial \mathrm{EP}(a)/\partial z_i| \le 4\sigma^2 / N, \quad |\partial^2 \mathrm{EP}(a)/\partial z_i^2| \le C\sqrt{\pi}\, \sigma^3 / (2N),$$

dove C è una costante assoluta. Questo bound, derivato dalla regola della catena applicata alla funzione caratteristica empirica e dalla limitatezza di $|e^{itX}| = 1$, garantisce che i gradienti di SIGReg siano uniformemente controllati per qualsiasi distribuzione degli embedding, il che è la condizione necessaria per l'incorporazione in un ciclo di ottimizzazione basato su discesa del gradiente stocastica. La complessità computazionale dell'intera procedura è O(N) in memoria e tempo — lineare nel numero di campioni — grazie all'operazione all_reduce sull'ECF che il calcolo distribuito supporta nativamente.

La domanda sulla sufficienza di M direzioni in K dimensioni è risolta dal Teorema 5 di LeJEPA, che usa il Lemma di Cramér-Wold sferico (la coincidenza delle proiezioni unidimensionali su tutte le direzioni è equivalente all'eguaglianza in distribuzione) e la teoria delle approssimazioni nelle varietà sferiche. Denotando la regolarità di Sobolev della densità p_θ come $p_\theta \in H^\alpha(\mathbb{R}^K)$, il bound sull'errore direzionale atteso soddisfa:

$$\mathbb{E}_a\left[\int_{\mathbb{R}} |\varphi_a(t) - \varphi_{\mathcal{N}}(t)|^2 \, dt\right] \le C(K, \alpha) \cdot |\mathbb{A}|^{-2\alpha/(K-1)} \cdot \int_{\mathbb{S}^{K-1}} \|\varphi_\cdot(r) - \varphi_{\mathcal{N}}(r)\|^2_{H^\alpha(\mathbb{S}^{K-1})} \, dr,$$

con $C(K, \alpha) = 2^{2\alpha}\, \pi^{(K-1)/2}\, \Gamma(\alpha + (K-1)/2) / ((K-1)\, \Gamma(\alpha)\, \Gamma((K-1)/2))$. Il bound decade come $|\mathbb{A}|^{-2\alpha/(K-1)}$, il che significa che $|\mathbb{A}| = O(K)$ direzioni sono sufficienti per un'approssimazione ε-accurata quando α è grande. La distribuzione gaussiana è infinitamente differenziabile ($\alpha = \infty$ in senso formale), e il fatto che SIGReg spinge p_θ verso la gaussiana significa che α cresce nel corso del training, accelerando la convergenza dell'approssimazione. Il SGD amplifica ulteriormente questo effetto: anche con $|\mathbb{A}|$ piccolo (nella pratica 16–256 direzioni per step), il ricampionamento delle direzioni ad ogni minibatch produce una copertura cumulativa che cresce linearmente con il tempo di training, compensando il bound limitato per ogni singolo step.

Per il sistema AMI, le implicazioni di questo risultato sono di due tipi. Il primo riguarda il training del Perception Module: l'aggiunta di SIGReg come termine di regolarizzazione alla loss di I-JEPA trasforma il Perception Module da un sistema che tende verso l'ottimalità in modo implicito e fragile (attraverso lo stop-gradient e l'EMA) in uno che è spinto esplicitamente verso la distribuzione provabilmente ottimale. Questo elimina la necessità di monitorare la qualità delle rappresentazioni attraverso probing supervisionato durante il training — una pratica costosa e dipendente da dataset etichettati — sostituendola con la loss di training stessa, che mostra correlazione di Spearman del 94–99% con l'accuratezza downstream grazie alla linearità del fronte di trade-off nel piano (SIGReg loss, prediction loss). Il secondo tipo di implicazione riguarda la modularità di AMI: se il Perception Module garantisce embedding gaussiani isotrópici, i moduli downstream (World

Model, Cost Module, Actor) possono essere progettati a partire da questa assunzione distributiva, semplificando le loro funzioni di perdita e le procedure di inizializzazione.

V.9bis — GMJE come implementazione concreta del Perception Module probabilistico

Le sezioni V.9 e V.10 hanno stabilito che LeJEPA e STP forniscono rispettivamente la caratterizzazione distributiva ottimale (gaussiana isotropica) e la regolarizzazione geometrica delle traiettorie (geodetiche) per il Perception Module di AMI. Il framework GMJE (Huang, 2026) completa questo quadro offrendo una terza componente finora mancante: un modello generativo esplicito della struttura di coppia (z_c, z_t) che consente sia il ragionamento probabilistico condizionale sia il campionamento generativo incondizionato nello spazio delle rappresentazioni.

Nell'architettura AMI, il Perception Module deve assolvere a due funzioni simultane: costruire rappresentazioni ŝ_t = E_θ(x_t) per il World Model (funzione discriminativa) e fornire una stima probabilistica dell'incertezza su ŝ_t data l'osservazione parziale o ambigua x_t (funzione epistemica). GMJE-MDN realizza naturalmente entrambe le funzioni: il predittore condizionale p(z_t | z_c) = Σ_k γ_k(z_c) $\mathcal{N}$(z_t | μ_{t|c,k}(z_c), Σ_{t|c,k}(z_c)) fornisce la distribuzione completa sulla regione mascherata, con l'entropia H[p(z_t | z_c)] = −Σ_k γ_k log γ_k + Σ_k γ_k H[$\mathcal{N}$_k] come stima calibrata dell'incertezza epistemica. Questo è il termine di epistemic cost IC_{epistemic} della sezione V.5, ora derivabile in forma chiusa dalla struttura GMM senza richiedere ensemble o dropout bayesiano.

La Legge della Varianza Totale (Appendice K.3 di Huang, 2026) decompone l'incertezza predittiva in due componenti interpretabili: Σ_total = Σ_k γ_k Σ_{t|c,k} + Σ_k γ_k(μ_{t|c,k} − μ_total)(μ_{t|c,k} − μ_total)^T. Il primo termine è l'incertezza aleatorica locale (rumore intrinseco di ciascun modo), il secondo è l'incertezza epistemica multimodale (varianza tra i centri dei modi). Questa decomposizione — assente in I-JEPA, LeJEPA e nei metodi contrastivi — permette al Cost Module di AMI di distinguere ambiguità fondamentale (regioni dove il mondo ha multiple evoluzioni plausibili) da incertezza dovuta a informazione insufficiente (regioni dove un'osservazione aggiuntiva risolverebbe l'ambiguità). Il primo tipo dovrebbe guidare una policy più conservativa; il secondo tipo dovrebbe incentivare l'esplorazione attiva — esattamente il trade-off che la sezione V.5 identifica come centrale per l'agente AMI.

Sul piano computazionale, la revisione delle stime di Compute di VI.12 beneficia di GMJE attraverso tre canali aggiuntivi rispetto a STP e LeJEPA. Primo: SMC-GMJE riduce la dimensione minima della memory bank necessaria, diminuendo il footprint della memoria durante il training. Secondo: la struttura MDN permette di estrarre l'incertezza senza forward pass aggiuntivi (zero costo extra rispetto alla predizione standard). Terzo: i K prototipi globali di GMJE-Proto fungono da semantic codebook per il transfer learning, riducendo il numero di

campioni necessari per l'adattamento a nuovi domini — un vantaggio diretto per gli scenari in-domain analizzati nella sezione VI.11.

V.10 — La Semantic Tube Prediction come minimalizzazione dell'azione lungo le geodetiche latenti

La connessione tra STP e il framework AMI emerge da una prospettiva che sarebbe rimasta oscura se ci si fosse limitati a leggere STP come un semplice miglioramento dell'efficienza dei dati per gli LLM. Huang, LeCun e Balestriero (2026) articolano esplicitamente la relazione con gli Energy-Based Models nella sezione di lavori correlati del loro articolo, affermando che STP estende la filosofia degli EBM dalla minimizzazione dell'energia in stati specifici alla minimizzazione dell'azione — l'integrale della lagrangiana lungo la traiettoria di generazione. Questa generalizzazione è concettualmente profonda e merita di essere sviluppata nel contesto di AMI con la cura che richiede.

Un EBM classico — e in particolare il modulo Cost Module di AMI nella sua forma più semplice — definisce una funzione di energia $F_\theta(\hat{s}, a)$ che assegna un costo scalare a ogni configurazione stato-azione. La pianificazione in AMI consiste nel trovare la sequenza di azioni $(a_1, \ldots, a_T)$ che minimizza il costo cumulativo:

$$\operatorname{argmin}_{a_1,\ldots,a_T} \Sigma_{t=1}^{T} F_\theta(\hat{s}_t, a_t),$$

dove $\hat{s}_{t+1} = WM_\theta(\hat{s}_t, a_t)$ è la transizione predetta dal World Model. Questo schema è una minimizzazione dell'energia in stati discreti: si assegna un costo a ogni configurazione e si cerca il percorso di costo minimo nello spazio degli stati.

STP introduce una prospettiva radicalmente diversa: invece di minimizzare la somma dei costi sugli stati visitati, si minimizza la geometria intrinseca del percorso stesso nello spazio latente. La perdita $\mathcal{L}_{STP} = 1 - \cos(h_t - h_r, h_r - h_s)$, dove $s < r < t$ sono indici di token scelti casualmente, misura quanto il vettore di transizione intermedio $(h_r - h_s)$ si discosta dalla direzione del vettore di transizione globale $(h_t - h_s)$. In termini geometrici differenziali, questo è una penalizzazione della curvatura locale del percorso nel spazio latente: un percorso con curvatura zero è una retta, e $\mathcal{L}_{STP} = 0$ se e solo se $h_r - h_s$ è parallelo a $h_t - h_s$ per ogni tripla $s < r < t$.

Il collegamento con il Principio di Minima Azione — richiamato esplicitamente nell'articolo STP — consente di interpretare questa penalizzazione all'interno del formalismo lagrangiano. Se si modella la traiettoria degli stati nascosti come un cammino in $\mathbb{R}^d$ parametrizzato dal tempo discreto t, la Lagrangiana del sistema è $L(h, \dot{h})$ con $\dot{h} = h_t - h_{t-1}$ la "velocità" discreta. Il Principio di Minima Azione afferma che il cammino fisicamente realizzato minimizza l'azione:

$$S[h] = \Sigma_{t=1}^{T} L(h_t, h_t - h_{t-1}).$$

Per la lagrangiana cinematica $L = (1/2)\,\|\dot{h}\|^2$, che corrisponde a una particella libera, le traiettorie che minimizzano $S[h]$ sono le geodesite — rette nello spazio euclideo. La Geodesica Hypothesis di STP afferma che le traiettorie degli stati nascosti error-free di un LLM convergente sono geodetiche nel senso che la varianza del manifold su cui si trovano rende localmente lineare ogni segmento di lunghezza $\leq \tau$. La perdita $\mathcal{L}_STP$ è quindi un regolarizzatore che forza le traiettorie degli stati nascosti durante l'addestramento ad avvicinarsi a quelle che un agente razionale con lagrangiana cinematica percorrerebbe spontaneamente.

L'interpretazione formale di questo parallelismo richiede la dimostrazione che il processo di generazione sequenziale di token è modellabile tramite un'ODE. La Proposizione 2.1 di STP stabilisce questo risultato: se $x_{\leq t} \in \mathbb{R}^{T \times d_model}$ è la sequenza di token come vettori di embedding, allora la dinamica di training converge a:

$$dx_{\leq t} = \hat{u} \circ f(x_{\leq t})\, dt,$$

dove f è la funzione del network convergente ed $\hat{u}$ è la funzione di unembedding. La dimostrazione procede definendo l'operatore di prefix-removal $\ominus$ tramite la funzione di lift $v(x, t) = [0, \ldots, 0, x]$ *{indice t+1}, 0, …, 0]* $\in \mathbb{R}^{T \times d_model}$*, e mostrando che* $x_{\leq t+1} \ominus x_{\leq t} = x_{\leq t+1} - x_{\leq t}$ sotto quella convenzione, trasformando l'equazione di ricorrenza $x_{\leq t+1} = x_{\leq t} \oplus \hat{u} \circ f(x_{\leq t})$ in un'equazione della forma $z_{t+1} - z_t = g(z_t, t)$ che ammette modellazione ODE. Il Teorema di Picard-Lindelöf garantisce l'unicità della soluzione: se $\hat{u} \circ f$ e le sue derivate parziali rispetto a $x_{\leq t}$ sono continue, l'ODE ammette un'unica soluzione per ogni condizione iniziale. In termini di generazione, questo significa che sequenze generate da prompt distinti (condizioni iniziali diverse) non possono mai intersecarsi — il che esclude teoricamente il mode collapse — e che la traiettoria future dello stato latente è completamente determinata dallo stato corrente.

A tempo di inferenza, tuttavia, l'errore residuo $\varepsilon_t = h_t - h_t$ *(dove ht =* $f(x_{\leq t})$ è la traiettoria ottimale) si accumula producendo un processo stocastico. Nell'approssimazione a larghezza infinita (Yang e Littwin, 2021), le pre-attivazioni di una rete neurale convergono a processi gaussiani, il che permette di modellare ε_t come rumore bianco gaussiano i.i.d. La somma cumulativa di questi errori, applicata al teorema di Donsker, porta alla Proposizione B.1 di STP:

$$dx_{\leq t} = \hat{u} \circ f(x_{\leq t})\, dt + \sigma_t\, dW_t,$$

ovvero la dinamica di inferenza è un'SDE con termine di drift deterministico e termine stocastico browniano. La Proposizione F.1 quantifica il danneggiamento:

$$\|h_t - h^*_t\|_2 \propto \sigma\sqrt{t},$$

il raggio del cono di incertezza cresce come la radice quadrata del tempo. La diminuzione di σ — che STP ottiene rendendo le traiettorie latenti più regolari e quindi riducendo la componente di rumore ε_t — si traduce direttamente in un

cono di inferenza più stretto, ovvero in una minore probabilità di mode collapse per sequenze lunghe.

Questa derivazione ha una connessione diretta con la V.5 della monografia sulla modellazione dell'incertezza in AMI. Si era osservato che il Cost Module di AMI potrebbe incorporare un termine di costo epistemico proporzionale all'entropia della distribuzione predittiva del World Model:

IC_epistemic(ŝ_t, a_t) = -λ · H(p_θ(ŝ_{t+1} | ŝ_t, a_t)).

L'analisi SDE di STP suggerisce un collegamento più diretto: se il Perception Module produce traiettorie latenti con deviazione $\sigma\sqrt{t}$ dalla traiettoria ottimale, e se il World Model usa queste traiettorie come input, allora l'entropia H(p_θ(ŝ_{t+1} | ŝ_t)) sarà proporzionale a $\sigma^2 t$ — il quadrato del raggio del cono di inferenza. Un Perception Module addestrato con $\mathcal{L}$_STP riduce σ, il che si traduce in un abbassamento dell'incertezza epistemica del World Model senza richiedere modifiche architetturali al World Model stesso. STP agisce dunque indirettamente sull'incertezza epistemica dell'intero sistema AMI attraverso la regolarizzazione della qualità delle rappresentazioni prodotte dal Perception Module.

Il Lemma di Straightening (Lemma 3.2 di STP) fornisce la certezza quantitativa del confinamento nella Semantic Tube. Si supponga che hs = h_*s* e *ht* = *h*_t (condizioni al contorno soddisfatte dai token speciali ‹before-bos› e ‹after-eos›), e che $\mathcal{L}$_STP ≤ ε per tutti gli r con s < r < t. Allora, per piccoli ε:

$\|(h_r - h_s)_{\perp h_t - h_s}\|_2 \leq \sqrt{(2\varepsilon)} \cdot \|h_r - h_s\|_2$.

La dimostrazione sfrutta la relazione geometrica θ' ≥ θ tra l'angolo θ' tra (h_r - h_s) e (h_t - h_r) — direttamente penalizzato da $\mathcal{L}$_STP — e l'angolo θ tra (h_r - h_s) e la direzione ottimale (h_*t* - *h*_s). Con $\mathcal{L}$_STP = 1 - cos θ' ≤ ε e l'approssimazione cos θ' ≈ 1 - θ'²/2 per piccoli θ':

$\theta'^2/2 \lesssim \varepsilon \Longrightarrow \theta' \lesssim \sqrt{(2\varepsilon)} \Longrightarrow \sin\theta' \lesssim \sqrt{(2\varepsilon)}$,

e quindi $\|(h_r - h_s)_{\perp}\|_2 = \|h_r - h_s\|_2 \cdot \sin\theta \leq \|h_r - h_s\|_2 \cdot \sin\theta' \lesssim \sqrt{(2\varepsilon)} \cdot \|h_r - h_s\|_2$. Il Teorema della Semantic Tube estende questo a una garanzia globale: se h* è localmente lineare (nel senso che $\|(h_r - h_s)_{\perp h_t - h_s}\|_2 \leq \varepsilon$ per ogni tripla con |t - s| ≤ τ) e $\mathcal{L}$_STP → 0, allora $\|h_r - h^*\|_2 \lesssim \varepsilon$. Il corollario casuale (Corollario 3.4) afferma che, per s < r < t scelti uniformemente a caso dalla distribuzione dell'indice, per ogni ε e δ:

$P(\|h_r - h^*\|_2 > \varepsilon + \delta) \to 0$ quando $\mathcal{L}$_STP → 0,

ovvero il confinamento avviene con alta probabilità anche per tripli scelti casualmente, il che corrisponde all'implementazione pratica dove s, r, t sono selezionati uniformemente da un minibatch.

La dimostrazione che l'identità (e non una proiezione appresa) è il predittore ottimale in STP (Previsione P5, validata dalla sezione 4.6) è particolarmente

rilevante nel contesto AMI. In I-JEPA, il predittore è un Transformer separato che trasforma la rappresentazione del contesto in quella del target; il suo ruolo è essenziale perché la struttura dell'informazione condizionale p*(z_T | z_C, m) può essere complessa e non catturabile da una funzione identità. In STP, il predittore è l'identità: la linearità locale della geodetica garantisce che (h_r - h_s) sia già un buon predittore di (h_t - h_r), senza necessità di trasformazione. Questo contrasto illumina la distinzione funzionale tra I-JEPA e STP nel contesto AMI: I-JEPA cattura le relazioni strutturali multi-view all'interno di un singolo input (relazioni spaziali tra regioni dell'immagine), mentre STP cattura le relazioni dinamiche lungo una sequenza temporale (la struttura geodetica del cammino nel manifold latente). Le due fonti di struttura sono ortogonali e complementari — la prima riguarda la geometria interna di uno stato latente, la seconda la geometria della sua evoluzione temporale — e la loro integrazione in un unico sistema corrisponde esattamente alla struttura duale del Perception Module più World Model in AMI.

V.11 — Polimorfismo e preservazione della diversità: le implicazioni per il Cost Module

Un risultato di STP che merita attenzione particolare nel contesto di AMI è quello sulla preservazione della diversità, documentato nella sezione 4.4 e caratterizzato come "polimorfismo" dall'analisi SVD delle rappresentazioni. Si è trovato che STP apprende simultaneamente pattern funzionalmente equivalenti ma statisticamente distinti — nel caso del dataset NL-RX-SYNTH, le varianti "." *e* ".*" delle espressioni regolari, con la prima 35 volte più frequente della seconda — raggiungendo un'accuratezza dell'88,5% sulla prima e del 68,0% sulla seconda, contro il 29,9% e 28,0% del fine-tuning standard.

L'analisi SVD della differenza Enc(Text) - Enc(Code) rivela la struttura del polimorfismo: senza normalizzazione, il profilo degli autovalori singolari di STP somiglia a quello del fine-tuning standard (varianza ampia sui vettori grezzi); con normalizzazione, somiglia a quello di LLM-JEPA (struttura semplice sulle direzioni). Questa duplicità indica che STP impone una struttura semplice alle direzioni normalizzate dei vettori di differenza (catturando le preferenze semantiche del creatore del dataset) lasciando però libertà di variazione nelle magnitudini grezze (catturando la diversità statistica del corpus).

Nel linguaggio degli EBM, questo polimorfismo corrisponde a una funzione di energia che ha bacini attrattori multipli di bassa energia, separati da barriere sufficientemente alte da evitare la convergenza a un unico stato. Il fatto che STP preservi questa struttura multi-basin — mentre il fine-tuning standard tende a collassare verso il bacino più frequente — è direttamente rilevante per il Cost Module di AMI: un Cost Module che mappa stati latenti prodotti da un Perception Module addestrato con $\mathcal{L}$_STP sarà esposto a una distribuzione di stati che preserva la diversità semantica del dominio, permettendo una pianificazione più ricca rispetto a una che operi su rappresentazioni collassate verso i pattern dominanti. La preservazione della diversità è il complemento funzionale dell'anti-mode-collapse garantito dall'ODE: l'ODE garantisce che traiettorie

distinte non convergano tra loro durante la generazione, mentre il polimorfismo garantisce che la distribuzione degli stati latenti mantenga sufficiente varietà statistica per supportare comportamenti adattativi multipli.

V.12 — Sintesi: il triangolo LeJEPA–STP–AMI e le questioni aperte

La discussione dei paragrafi V.9–V.11 consente ora di tracciare con precisione il triangolo di connessioni tra LeJEPA, STP e il framework AMI, identificando non solo le conferme reciproche ma anche le tensioni residue che rimangono da risolvere.

La convergenza è trilatera. LeJEPA dimostra che gli embedding ottimali per compiti downstream arbitrari sono gaussiani isotrópici, il che fornisce la giustificazione teorica per la struttura del Perception Module in AMI. STP dimostra che le traiettorie latenti ottimali per la generazione sequenziale sono geodetiche — rette sul manifold semantico — il che fornisce la giustificazione teorica per la struttura del World Model in AMI: le transizioni di stato latente $\hat{s}_{t+1} = WM_\theta(\hat{s}_t, a_t)$ dovrebbero, in un sistema ben condizionato, essere localmente lineari nelle rappresentazioni prodotte dal Perception Module. Il framework AMI, a sua volta, fornisce il contesto architetturale in cui sia LeJEPA sia STP trovano il loro pieno significato funzionale: non sono migliorie isolate di modelli esistenti ma componenti di un sistema integrato dove la qualità distributiva delle rappresentazioni (LeJEPA) e la regolarità geometrica delle loro traiettorie (STP) cooperano per produrre un agente capace di pianificazione a lungo termine.

Le tensioni residue sono di natura matematica e pratica. Sul piano matematico, la Geodesica Hypothesis di STP assume che il manifold su cui si trovano le traiettorie latenti abbia struttura liscia, il che è garantito solo asintoticamente per network di larghezza infinita (nel limite Neural Tangent Kernel studiato da Yang e Littwin). Per reti di larghezza finita — come tutti i modelli pratici — la curvatura del manifold è non nulla e variabile, il che spiega empiricamente perché $\lambda \approx 0.01$ sia il valore ottimale di compromesso: un λ troppo grande impone linearità dove la geodetica è intrinsecamente curvilinea, producendo una penalizzazione impropria. L'analisi di LeJEPA sull'ottimalità della gaussiana isotrópica assume, d'altro canto, che la varianza degli embedding sia controllata da un vincolo scalare fisso, mentre in pratica la varianza cambia durante il training e la costante $\varkappa$ deve essere mantenuta attraverso meccanismi di normalizzazione. La combinazione di SIGReg (che controlla la distribuzione marginale) e $\mathcal{L}_{STP}$ (che controlla la geometria delle traiettorie) non è ancora stata studiata come sistema integrato per grandi modelli di linguaggio o di visione, e la domanda se i loro effetti siano additivi, sinergici o potenzialmente antagonisti rimane aperta alla ricerca sperimentale.

Sul piano pratico, l'implementazione congiunta di LeJEPA e STP in un Perception Module per AMI richiederebbe di risolvere questioni di scala computazionale: SIGReg è O(N) in memoria e tempo, ma richiede il campionamento di M direzioni ad ogni step e la valutazione della statistica di Epps-Pulley su M proiezioni del minibatch; $\mathcal{L}_{STP}$ richiede invece la selezione di triple casuali (s, r, t) dall'output

dell'ultimo layer dell'LLM, un'operazione O(1) rispetto al forward pass. La compatibilità architetturale dei due termini con l'obiettivo predittivo di I-JEPA (la loss di predizione nello spazio latente delle patch mascherate) non è stata ancora studiata, e la domanda su come bilanciare tre contributi distinti alla funzione obiettivo — $\mathcal{L}$_pred (predizione JEPA), SIGReg (regolarizzazione distributiva), $\mathcal{L}$_STP (regolarizzazione geodetica) — è aperta. Un'ipotesi di lavoro teoricamente motivata è che SIGReg operi principalmente sullo spazio degli embedding all'interno di ogni input (garantendo isotropia marginale) mentre $\mathcal{L}$_STP operi sulla geometria delle sequenze di embedding attraverso il tempo (garantendo linearità delle traiettorie), con i due vincoli ortogonali sullo spazio dei gradi di libertà del network: il primo sulla distribuzione delle rappresentazioni statiche, il secondo sulla dinamica delle loro variazioni.

Capitolo VI — Fattibilità, Scala e Sostenibilità: Il Cammino Computazionale verso l'AGI

VI.1 — Le Scaling Laws come bussola empirica

Il progresso sperimentale nell'intelligenza artificiale degli ultimi anni è stato guidato da una scoperta empirica di portata straordinaria: le prestazioni dei modelli di linguaggio e visione migliorano in modo prevedibile e regolare al crescere della scala di addestramento, seguendo leggi di potenza (power laws) che si estendono su molti ordini di grandezza. Questa regolarità, nota come Scaling Laws, è stata documentata sistematicamente da Kaplan et al. (2020) per i modelli linguistici e da successivi studi per i modelli multimodali e visivi.

Le Scaling Laws sono, in questo senso, uno dei più sorprendenti fenomeni empirici dell'intera storia delle scienze dell'informazione: la regolarità con cui le prestazioni dei modelli migliorano seguendo leggi di potenza su molti ordini di grandezza di compute, parametri e dati suggerisce che esiste una struttura profonda nell'apprendimento automatico — una sorta di «meccanica statistica» dell'intelligenza artificiale — che non è ancora compresa a livello teorico ma che si manifesta empiricamente con affidabilità notevole. Comprendere le origini di queste leggi — perché l'esponente α_D sia approssimativamente 0.095 e non 0.05 o 0.5 — è un problema aperto di notevole importanza sia per la fisica dell'apprendimento che per la pianificazione delle risorse computazionali per i sistemi futuri.

Le violazioni delle Scaling Laws documentate dalla Semantic Tube Prediction sono, in questo contesto, altrettanto sorprendenti quanto le leggi stesse: se queste leggi riflettono proprietà profonde dell'apprendimento, una loro violazione non è un fatto tecnico minore ma un segnale che la nostra comprensione del processo di apprendimento è incompleta in modo sistematico. STP suggerisce che le Scaling Laws caratterizzano una famiglia specifica di obiettivi di apprendimento — quelli definiti nello spazio delle osservazioni discrete — e che obiettivi con struttura

geometrica diversa possono operare in un regime di scaling qualitativamente distinto.

La forma matematica delle Scaling Laws per i modelli linguistici (Kaplan-style) è:

$$L(N, D, C) \approx (N_c / N)^{(\alpha_N)} + (D_c / D)^{(\alpha_D)} + L_\infty$$

dove L è la cross-entropy loss di test, N è il numero di parametri del modello, D è il numero di token di addestramento, $C \approx 6 \cdot N \cdot D$ è il compute totale in FLOPs (con il fattore 6 che approssima il costo di un forward pass e un backward pass attraverso un transformer di N parametri su D token), N_c e D_c sono costanti critiche che caratterizzano il punto di crossover, $\alpha_N \approx 0.076$ e $\alpha_D \approx 0.095$ sono gli esponenti della legge di potenza, e L_∞ è il limite inferiore della loss (l'entropia irriducibile del dominio).

La versione rivista di Chinchilla (Hoffmann et al., 2022) chiarisce che Kaplan et al. avevano usato modelli troppo piccoli relativamente ai dati, e che il rapporto ottimale tra parametri e token di addestramento è approssimativamente:

$$D_{opt} \approx 20 \cdot N$$

ovvero, per un modello di N parametri, il numero ottimale di token di addestramento (in termini di minimizzazione della loss per un budget di compute fisso) è circa 20N. Questa relazione è nota come "Chinchilla scaling" e ha ridefinito le pratiche di addestramento dei modelli linguistici: GPT-4 con ~1.8T parametri richiederebbe circa 36T token per un training Chinchilla-ottimale, un dataset di dimensioni simili all'intero testo disponibile su internet.

Per i modelli di visione preaddestrati con obiettivi autosupervisionati come I-JEPA, le scaling laws sono meno ben caratterizzate, ma i risultati disponibili suggeriscono esponenti simili:

$$L_{vision}(N, D) \approx (N_c / N)^{(\alpha_N)} + (D_c / D)^{(\alpha_D)}$$

con $\alpha_N \approx 0.07–0.10$ e $\alpha_D \approx 0.10–0.15$, valori lievemente più grandi di quelli per il linguaggio, il che suggerisce che i modelli visivi richiedono relativamente meno dati per parametro rispetto ai modelli linguistici per saturare le loro capacità.

VI.2 — Stima del compute necessario per sistemi AMI

Una stima anche approssimativa del compute necessario per addestrare un sistema AMI completo è utile non per fare previsioni precise — che sarebbero inevitabilmente sbagliate data l'evoluzione rapida del campo — ma per comprendere gli ordini di grandezza e identificare i colli di bottiglia.

Un sistema AMI minimale funzionante richiederebbe, secondo la visione di LeCun, i seguenti componenti di base:

Il Perception Module (I-JEPA) a scala sufficiente per ragionamento semantico complesso richiede presumibilmente un ViT con 1–10 miliardi di parametri

(ordine di grandezza di ViT-22B di Google, il più grande Vision Transformer addestrato fino al 2024) su un dataset di immagini/video di 100B–1T token visivi. Il compute per questo modulo, usando le Scaling Laws visive, è dell'ordine di:

$C_{perception} \approx 6 \cdot N_{perception} \cdot D_{perception} \approx 6 \cdot 10^{9} \cdot 10^{12} = 6 \times 10^{21}$ FLOPs

che corrisponde a circa 6 ZetaFLOPs, o equivalentemente a circa 10^{4}–10^{5} GPU-days su A100 (che esegue ~3×10^{14} FLOPs/s in half precision).

Il World Model che opera sullo spazio latente del Perception Module e deve modellare dinamiche complesse in ambienti realistici richiederebbe una scala comparabile ai modelli linguistici di punta — ordine di grandezza 10^{10}–10^{11} parametri addestrati su traiettorie di lunghezza 10^{7}–10^{9} step, con costo:

$C_{worldmodel} \approx 6 \cdot 10^{10} \cdot 10^{9} \cdot T_{context} \approx 6 \times 10^{23}$ FLOPs

dove $T_{context}$ è la lunghezza del contesto temporale (tipicamente 10^{2}–10^{3} step per sequenze di azione-stato). Questo è un ordine di grandezza di 0.6 YottaFLOPs, circa 10^{6} GPU-days su A100 — sostanzialmente oltre il compute degli attuali run di addestramento più grandi (GPT-4 è stimato a ~10^{23}–10^{24} FLOPs).

Questi numeri, pur essendo molto approssimati, hanno una conseguenza chiara: un sistema AMI a piena scala non è addestrabile oggi con le risorse computazionali disponibili, e forse non lo sarà per un decennio. Questo non implica che la ricerca debba fermarsi — al contrario, la sfida computazionale è uno stimolo per lo sviluppo di algoritmi più efficienti — ma impone un ridimensionamento delle ambizioni a breve termine.

VI.3 — Efficienza algoritmica e la legge di Moore computazionale

La storia dell'intelligenza artificiale è segnata non solo dalla crescita del compute disponibile (la legge di Moore hardware) ma anche da quella dell'efficienza algoritmica — la riduzione del compute necessario per raggiungere una certa prestazione grazie a migliori algoritmi. OpenAI ha documentato che tra il 2012 e il 2021 il compute necessario per addestrare un sistema al livello di ImageNet (top-5 accuracy su ImageNet) si è ridotto di un fattore ~44x grazie a miglioramenti algoritmici, indipendentemente dalla crescita dell'hardware. Questo tasso di miglioramento algoritmico — circa 2x ogni due anni — è comparabile alla legge di Moore hardware nella sua versione classica.

Prima di enumerare le direzioni di miglioramento algoritmico specifiche per JEPA e World Models, è utile quantificare il tasso storico di miglioramento dell'efficienza algoritmica e confrontarlo con quello dell'efficienza hardware, poiché i due trend hanno implicazioni diverse per il percorso verso l'AGI.

Hernandez e Brown (2020), analizzando sistematicamente i progressi nell'addestramento di classificatori su ImageNet dal 2012 al 2020, hanno stimato che il compute necessario per raggiungere un dato livello di prestazione si è ridotto

di circa 44× in otto anni — un raddoppio circa ogni undici mesi. Nello stesso periodo, l'efficienza dell'hardware (FLOP/s per dollaro) è migliorata di circa 2.5× negli stessi otto anni, un raddoppio ogni tre anni circa. L'efficienza algoritmica ha quindi migliorato circa tre volte più velocemente dell'efficienza hardware nel decennio 2012-2020.

Questo dato sorprendente ha implicazioni cruciali per la pianificazione: se il trend storico continua, investire risorse nella ricerca algoritmica produce ritorni computazionalmente superiori all'investimento in hardware. Un'architettura che richiede 100× meno compute per raggiungere lo stesso livello di prestazione non è soltanto più economica: è accessibile a un numero enormemente maggiore di ricercatori e laboratori, democratizzando la ricerca e accelerando il ciclo di scoperta.

JEPA e i World Models si collocano in questa prospettiva come candidati ad alta efficienza algoritmica potenziale: la predizione latente, come dimostrato empiricamente da LeJEPA e da STP, può ridurre significativamente il numero di dati necessari rispetto agli obiettivi standard, il che si traduce direttamente in riduzione del compute a parità di prestazione. Il fattore $\varrho_STP \approx 16$ documentato per la Semantic Tube Prediction è di un ordine di grandezza simile ai miglioramenti algoritmici decennali documentati da Hernandez e Brown — suggerendo che un singolo cambio di obiettivo di addestramento può produrre un salto di efficienza equivalente a diversi anni di progresso hardware.

Per i modelli di tipo JEPA e World Model, alcune direzioni di miglioramento algoritmico sono già visibili:

Per i modelli di tipo JEPA e World Model, alcune direzioni di miglioramento algoritmico sono già visibili:

La prima è l'**apprendimento curriculum** con esposizione adattiva: anziché campionare casualmente i dati di addestramento, si usano strategie di sampling che espongono il modello prima ai concetti più frequenti e poi ai casi rari, riducendo il numero totale di passi necessari per convergenza. Studi su LLM mostrano riduzioni del compute fino al 30–50% con curriculum learning appropriato.

La seconda è la **distillazione da modelli grandi a modelli piccoli**: una volta addestrato un modello grande, si addestra un modello più piccolo per imitarne le rappresentazioni o le predizioni. Questo permette di usare il modello piccolo per inferenza con costo molto ridotto. La distillazione può ridurre il numero di parametri del fattore 10–100 con perdita di prestazioni limitata (tipicamente 1–5% su benchmark standard).

La terza è l'**addestramento a precisione ridotta con quantizzazione**: i modelli moderni vengono addestrati in bfloat16 (16 bit) e possono essere eseguiti in int8 o int4 per l'inferenza senza perdita significativa di qualità, riducendo il costo computazionale per l'inferenza di un fattore 2–4 rispetto al float32.

La quarta, specifica per I-JEPA e i World Models, è la **parametrizzazione condivisa tra encoder e World Model**: se l'encoder percettivo e il modello dinamico condividono i layer inferiori (che catturano feature generiche) e differiscono solo nei layer superiori (dove si specializzano per percezione e predizione rispettivamente), si può ridurre il parametro totale del sistema senza sacrificare capacità. Questo principio — analogo al multi-task learning nei transformer linguistici — è stato applicato in modo parziale in V-JEPA ma non ancora sistematicamente in architetture AMI complete.

VI.4 — Le leggi di scala per i World Models

Una questione empirica non ancora risolta con soddisfazione è: i World Models seguono le stesse scaling laws dei modelli linguistici? E se sì, quali sono gli esponenti?

L'analogia strutturale è suggestiva: sia i LLM che i World Models sono transformer autoregressivi addestrati su sequenze — di token testuali nel primo caso, di token (stato, azione) nel secondo — e la complessità del dominio è in entrambi i casi aperta. Se la legge di potenza $L(N, D) \approx N^{(-\alpha)}$ con $\alpha \approx 0.07$ si applicasse uniformemente ai World Models, allora per ridurre la loss di un fattore 2 sarebbe necessario aumentare il numero di parametri di un fattore $2^{(1/0.07)} \approx 10^4$, ovvero quattro ordini di grandezza. Questo suggerisce che i guadagni marginali del solo scaling parametrico dei World Models sono relativamente modesti, il che motiva la ricerca di inductive biases architetturali specifici per la fisica e la causalità.

I dati disponibili per DreamerV3 mostrano miglioramenti quasi monotoni delle prestazioni su Atari all'aumentare della scala del modello (da 10M a 200M parametri), ma non è ancora chiaro se la tendenza si estenda a modelli più grandi o se vi sia un punto di saturazione. Per GAIA-1, i 9B parametri producono video di guida più realistici di versioni più piccole, ma la valutazione quantitativa della qualità dei World Models visivi — a differenza dei LLM dove la perplexity su corpus di test è una metrica pulita — rimane metodologicamente complicata.

Un ulteriore punto delicato riguarda la **data efficiency** dei World Models rispetto ai LLM. Un agente che impara un World Model attraverso l'interazione con il mondo reale non può permettersi di processare 10^{12} token di esperienza: il numero di ore di interazione motoria di un bambino umano dal nascere all'adulto è dell'ordine di 10^5–10^6 ore, corrispondente a circa 10^{10}–10^{11} frame video (a 30fps) o 10^9–10^{10} step di controllo motorio. Se le scaling laws si applicano con gli stessi esponenti, un World Model che impara dall'interazione reale dovrebbe essere molto più piccolo di uno addestrato su dati passivi di web-scale — il che suggerisce che l'interazione embodied e il feedback di errore predittivo in tempo reale possono essere forme di supervisione molto più efficienti dei dati passivi.

VI.5 — Sostenibilità energetica: la sfida silenziosa

Il dibattito pubblico sull'intelligenza artificiale è spesso dominato da preoccupazioni di natura etica e sociale — bias, disinformazione, impatto sul

lavoro — che sono certamente rilevanti, ma che tendono a oscurare una sfida di carattere fisico e infrastrutturale che potrebbe rivelarsi il vincolo più stringente al percorso verso l'AGI: il consumo energetico.

Addestrare GPT-4 ha richiesto, secondo stime pubblicamente disponibili, circa 10^{23}–10^{24} FLOPs. Con un'efficienza media degli acceleratori hardware di ~10^{14} FLOP/s/W (circa quella degli A100 in condizioni operative tipiche di cluster), il consumo energetico per questo addestramento è stato dell'ordine di:

$E_training \approx 10^{23}$ FLOP / 10^{14} FLOP/s/W ≈ 10^{9} W·s ≈ 277 MWh

Questo valore equivale al consumo annuale di circa 25 abitazioni europee medie — non trascurabile, ma non ancora catastrofico. Il problema non è il singolo run di addestramento ma la traiettoria: se il compute degli addestramenti cresce di 10x ogni due anni (come osservato empiricamente tra il 2012 e il 2024), e se questa traiettoria continua, il consumo energetico per singolo run raggiungerà l'ordine di grandezza del TWh entro il 2030 e del PetaWh entro il 2040. Quest'ultimo valore corrisponde a circa la metà della produzione elettrica annuale mondiale — chiaramente insostenibile.

Prima di descrivere le tre vie principali, è utile essere più precisi sulla natura del problema energetico, che spesso viene discusso in termini eccessivamente astratti. Il consumo energetico di un sistema di apprendimento automatico ha due componenti principali: il costo dell'addestramento (un costo "one-shot" che viene pagato una volta per produrre il modello) e il costo dell'inferenza (un costo "ricorrente" che viene pagato ad ogni utilizzo del modello).

Per i grandi modelli attuali, il costo dell'addestramento è tipicamente dominante durante lo sviluppo ma trascurabile rispetto al costo totale dell'inferenza su scale di deployment massivo. GPT-4 ha richiesto un addestramento stimato in 200-300 MWh — il consumo annuale di circa 20-30 abitazioni europee. Ma se GPT-4 viene interrogato da 100 milioni di utenti per una media di 10 query al giorno, ciascuna della durata di un secondo a una potenza stimata di 0.001 kWh, il costo di inferenza annuale ammonta a circa 365.000 MWh — mille volte il costo dell'addestramento. Il dibattito sul consumo energetico dell'IA che si concentra sull'addestramento sta guardando la punta dell'iceberg.

Questa osservazione ha implicazioni dirette per la valutazione del framework AMI. Un sistema AMI che richiede 10× meno compute per l'addestramento grazie all'efficienza di JEPA e STP — ma che impiega un World Model con inferenza lenta per ogni decisione — potrebbe avere un consumo energetico a lungo termine superiore a un sistema basato su policy reattive veloci. L'ottimizzazione del consumo energetico per sistemi AGI richiederebbe quindi di considerare separatamente i costi di addestramento e di inferenza, e di progettare architetture che bilancino i due in modo appropriato per il contesto di deployment.

Esistono tre vie principali per affrontare questo problema:

Esistono tre vie principali per affrontare questo problema:

La prima è l'**efficienza algoritmica** già discussa: algoritmi più efficienti riducono il compute necessario per unità di capacità, potenzialmente annullando la crescita del consumo energetico anche in presenza di modelli più grandi. I progressi in questo ambito sono stati storicamente più veloci di quanto ci si aspettasse, ma non è garantito che continuino allo stesso ritmo.

La seconda è l'**efficienza hardware**: i chip di nuova generazione (NVIDIA H100, Google TPU v5, chip custom di Amazon, Microsoft, Meta) mostrano miglioramenti di efficienza energetica significativi rispetto alla generazione precedente, tipicamente nell'intervallo 2–5x per generazione. Se questo tasso di miglioramento continua, il consumo energetico per unità di capacità computazionale si dimezzerebbe ogni tre-quattro anni.

La terza via — meno discussa ma potenzialmente la più impattante a lungo termine — è lo **spostamento verso paradigmi di apprendimento radicalmente più efficienti** dal punto di vista dei dati. Se un sistema AMI con World Model interattivo può imparare da 10^9 step di interazione piuttosto che da 10^{13} token di dati passivi — come la biologia suggerisce essere possibile — allora il compute totale necessario si riduce di diversi ordini di grandezza anche senza miglioramenti algoritmici o hardware. Questa è, in ultima analisi, la promessa più audace del framework AMI: che l'apprendimento grounded attraverso l'interazione sia intrinsecamente più efficiente dell'apprendimento passivo da dati web-scale.

VI.6 — La questione dei dati: scarsità, qualità e sinteticità

Il bottleneck del compute è correlato ma distinto dal bottleneck dei dati. I modelli di linguaggio e visione moderni sono addestrati su dati di qualità molto variabile estratti dal web, e vi è una crescente evidenza che la qualità dei dati sia tanto importante quanto la quantità. Lavori come Phi-1 di Microsoft (2023) e LIMA di Meta (2023) mostrano che modelli di dimensione relativamente contenuta (1.3B–65B parametri) addestrati su dati di alta qualità selezionati con cura possono competere con modelli molto più grandi addestrati su dati non curati.

Per i World Models e i sistemi AMI, il problema dei dati ha una dimensione aggiuntiva: i dati devono essere non solo linguisticamente o visivamente ricchi ma **strutturalmente causali** — ovvero devono contenere sequenze di azioni e conseguenze che permettano al sistema di inferire le leggi fisiche e causali del mondo. Questo tipo di dati è intrinsecamente più raro dei dati visivi passivi: un video di YouTube mostra il mondo ma non registra le intenzioni e le decisioni di chi agisce in esso; un dataset di gioco robotico registra le azioni ma in ambienti molto ristretti rispetto alla ricchezza del mondo reale.

Una direzione promettente è la **generazione di dati sintetici** attraverso simulatori fisici: ambienti come IsaacGym di NVIDIA, Mujoco, o il simulatore di Waymo permettono di generare traiettorie di interazione in quantità arbitraria a costo computazionale relativamente basso. Il problema del sim-to-real gap — la

difficoltà di trasferire modelli addestrati in simulazione al mondo reale — è tuttora una sfida aperta, ma studi recenti su robot policy come RT-2, SayCan e OpenVLA mostrano che con Domain Randomization e tecniche di adattamento del dominio, il gap può essere ridotto a livelli operativamente accettabili.

La questione dei dati porta naturalmente a una considerazione sull'architettura del sistema di apprendimento: un sistema che impara dal mondo reale attraverso l'interazione — come un bambino o un robot — non ha accesso a un dataset fisso di traiettorie ma genera continuamente nuovi dati attraverso le proprie azioni. Questo schema di **continual learning with exploration** è fondamentalmente diverso dall'apprendimento batch standard su un dataset fisso, e richiede architetture che possano aggiornarsi continuamente senza catastrophic forgetting e che possano usare l'incertezza predittiva come segnale per guidare l'esplorazione verso regioni informativamente ricche.

VI.7 — Stime temporali e scenari per l'AGI

Qualsiasi discussione seria sull'intelligenza artificiale generale non può evitare la domanda sull'orizzonte temporale, pur riconoscendo l'estrema incertezza di qualsiasi stima. Si propongono qui tre scenari distinti basati su assunzioni diverse sull'evoluzione del campo.

Lo **scenario di continuazione tendenziale** assume che: le scaling laws per LLM e World Models continuino a reggere su almeno altri due-tre ordini di grandezza di parametri e compute; che i miglioramenti algoritmici continuino al tasso storico di circa 2x ogni due anni; che i chip specializzati continuino a migliorare l'efficienza energetica di 2–3x per generazione. In questo scenario, sistemi con capacità generali comparabili a quelle umane in domini limitati (navigazione, manipolazione, reasoning quotidiano) potrebbero essere realizzabili nell'arco di 10–20 anni, con AGI nel senso pieno entro 30–50 anni.

Lo **scenario ottimistico** assume breakthrough algoritmici di tipo qualitativo — ad esempio, la scoperta di meccanismi di apprendimento radicalmente più efficienti analoghi a ciò che il cervello umano realizza con ~20W di potenza — che riducono il compute necessario per unità di capacità di diversi ordini di grandezza rispetto alla traiettoria attuale. In questo scenario, sistemi con capacità generali comparabili all'uomo potrebbero emergere nell'arco di 10–15 anni.

Lo **scenario pessimistico** assume che vi siano barriere di principio al scaling degli approcci attuali — ad esempio, che la composizionalità sistematica, il ragionamento causale o il grounding nel mondo reale richiedano cambiamenti architetturali fondamentali che non sono ancora stati identificati — e che queste barriere blocchino i progressi verso l'AGI per decenni. In questo scenario, i sistemi come AMI/JEPA/DreamerV3 produrranno agenti capaci in domini specifici ma non AGI nel senso pieno, almeno non nei prossimi 50–100 anni.

La posizione di LeCun stesso è interessante: egli è storicamente scettico rispetto alle previsioni ottimistiche a breve termine sull'AGI (ha sostenuto pubblicamente

che i LLM non possono essere la via verso l'AGI per ragioni di principio legate all'assenza di comprensione causale e grounding) ma è ottimista sulla fattibilità a lungo termine della via che propone. La sua stima informale è che un sistema AMI funzionante a livello di animale (non umano) potrebbe essere realizzabile nell'arco di 5–10 anni di ricerca focalizzata, con progressi verso l'intelligenza umana che richiederebbero decenni aggiuntivi.

VI.8 — Il problema della verifica: come sappiamo quando abbiamo raggiunto l'AGI

Una questione metodologica fondamentale che deve accompagnare qualsiasi discussione sull'AGI è: quali sono i criteri di verifica? Come si determina se un sistema ha raggiunto l'intelligenza generale?

Il Test di Turing — rispondere a domande in modo indistinguibile da un umano in una chat testuale — è chiaramente insufficiente alla luce dei LLM moderni, che lo superano in molti contesti pur essendo palesemente privi di comprensione nel senso rilevante. Benchmark come BIG-Bench, MMLU, HumanEval misurano capacità specifiche ma non generalità. L'ARC-AGI di François Chollet (2023) propone test di ragionamento astratto che valutano la capacità di generalizzare a nuovi pattern da pochi esempi, ma anche questo è un benchmark limitato che misura un aspetto della generalità piuttosto che la totalità.

Una proposta più robusta — sebbene difficile da operazionalizzare — è quella di valutare un sistema AGI rispetto alla sua capacità di apprendere nuovi compiti in modo efficiente (sample efficiency), di generalizzare a distribuzioni di dati significativamente diverse da quelle di training (out-of-distribution generalization), e di pianificare su orizzonti temporali lunghi in presenza di incertezza. Questi tre criteri corrispondono esattamente alle capacità che il framework AMI mira a sviluppare, il che rende AMI non solo una proposta architetturale ma anche, implicitamente, una teoria di cosa significhi intelligenza generale nel senso artificiale.

Il fatto che non esista ancora consensus sulla definizione operativa di AGI è esso stesso rilevante per la valutazione delle proposte come AMI: è possibile che sistemi costruiti secondo il framework di LeCun soddisfino alcune delle definizioni possibili di AGI molto prima di soddisfarne altre, e che il dibattito su "cosa conta come AGI" continui ben dopo che i sistemi in questione siano stati costruiti. Questa ambiguità definitoria non è solo accademica: ha conseguenze pratiche per la governance, la regolamentazione e la gestione dei rischi associati ai sistemi di IA avanzati — temi che vengono affrontati nell'ultimo capitolo della presente monografia.

VI.8.A — Considerazioni etiche e sociali sull'impatto dell'AGI computazionalmente intensiva

La sezione precedente ha discusso la sostenibilità energetica in termini puramente fisici e computazionali. È però necessario affrontare anche la dimensione

socio-economica del problema, che è complementare a quella tecnica e altrettanto urgente per la governance dell'intelligenza artificiale avanzata.

Il consumo computazionale dei sistemi di apprendimento automatico di grande scala non è distribuito uniformemente. Pochi laboratori — principalmente Meta, Google, Microsoft/OpenAI, Anthropic — hanno accesso alle risorse computazionali necessarie per addestrare modelli della scala richiesta per AMI. Questo squilibrio ha conseguenze per la diversità della ricerca, per la competizione tecnologica internazionale, e per la governance dell'intelligenza artificiale: se le tecnologie AGI più avanzate sono sviluppate e controllate da un numero ristretto di attori privati, le decisioni su come queste tecnologie vengono usate, distribuite e governate sfuggono al controllo democratico.

I risultati di LeJEPA e STP hanno potenzialmente conseguenze positive su questa dimensione: se il compute necessario per il Perception Module di AMI può essere ridotto da 10^{23} a 10^{22} FLOP, il range di attori capaci di sviluppare ricerca competitiva in questo campo si espande significativamente. Un laboratorio universitario con accesso a 100 GPU può esplorare varianti architetturali che con le risorse attuali richiederebbero un datacenter. Questa democratizzazione del compute non garantisce una distribuzione equa dei benefici dell'AGI, ma è una precondizione necessaria per essa.

Parallelamente, l'efficienza dell'addestramento rende più urgente la questione della sicurezza: se i sistemi AGI diventano più accessibili economicamente, il rischio che vengano sviluppati da attori con obiettivi disallineati aumenta. L'espansione della platea dei costruttori richiede una corrispondente espansione della platea dei ricercatori di sicurezza, dei regolatori e dei supervisori indipendenti. Queste considerazioni non appartengono al Capitolo VI della monografia in modo accidentale: la sostenibilità computazionale e la sostenibilità sociale dell'intelligenza artificiale generale sono problemi interconnessi che richiedono soluzioni integrate.

VI.9 — Revisione delle stime di compute alla luce dell'efficienza dei dati di STP e LeJEPA

Le stime di compute presentate nelle sezioni VI.1–VI.2 si fondano sulle Scaling Laws in forma Chinchilla, che prescrivono un rapporto ottimale tra parametri e token di addestramento pari a $D_{opt} \approx 20 \cdot N$ e predicono una dipendenza della cross-entropy loss dalla quantità di dati proporzionale a $(D_c/D)^{\alpha_D}$ con $\alpha_D \approx 0.095$. Un corollario implicito di queste leggi — mai esplicitato come assioma ma tacitamente assunto come limite fisico — è che la perdita è una funzione strettamente decrescente della quantità di dati per qualsiasi fissata architettura e obiettivo di addestramento. I risultati di Semantic Tube Prediction sfidano questo corollario in modo empiricamente diretto e teoricamente giustificato, con conseguenze quantificabili sulle stime di compute per AMI.

Il risultato empirico chiave è che STP permette a Llama-3 1B, 3B e 8B di raggiungere l'accuratezza del fine-tuning standard con soli D/16 dati di

addestramento sul dataset NL-RX-SYNTH, con degradazione trascurabile fino a D/8 e parità con il fine-tuning standard a D/16. In termini della parametrizzazione di Chinchilla, questo equivale ad affermare che l'obiettivo composto $\mathcal{L} = \mathcal{L}_NTP + \lambda \cdot \mathcal{L}_STP$ abbassa il valore critico D_c nella Scaling Law rispetto all'obiettivo $\mathcal{L}_NTP$ puro: lo stesso modello di N parametri raggiunge la medesima loss con D/16 dati invece di D. Formalmente, se si denota con $L^*(N, D, \mathcal{O})$ la loss attribuibile a N parametri, D token e obiettivo $\mathcal{O}$, le stime implicite nelle sezioni precedenti assumevano

$$L^*(N, D, \mathcal{L}_NTP) = (N_c/N)^{\alpha_N} + (D_c/D)^{\alpha_D} + L_\infty,$$

mentre i dati empirici di STP suggeriscono

$$L^*(N, D, \mathcal{L}_NTP + \lambda\, \mathcal{L}_STP) \approx L^*(N, D \cdot \varrho_STP, \mathcal{L}_NTP),$$

con $\varrho_STP \approx 16$ un fattore di amplificazione dell'efficienza dei dati. Il meccanismo teorico è fornito dal Lemma H.1 di STP (dimostrato nell'appendice H del paper): modellando il processo di addestramento come estrazione di informazione su un token target Y da m stati nascosti $\{X_i\}_{i=1}^m$ condizionalmente i.i.d. dato Y, si ha

$$H(Y \mid X^m) \geq H(Y) - m \cdot I(Y; X),$$

dove I(Y; X) è l'informazione mutua tra il token target e un singolo stato nascosto. Sotto l'approssimazione del canale gaussiano additivo $X = Z + N$ con $SNR = \mathbb{E}[\|Z\|^2] / \mathbb{E}[\|N\|^2]$ e la formula di capacità di Shannon

$$I(X; Y) = (1/2) \log(1 + SNR),$$

il numero minimo di campioni per raggiungere $H(Y \mid X^m) \leq \varepsilon$ soddisfa:

$$m \geq (H(Y) - \varepsilon) / ((1/2) \log(1 + SNR)).$$

Questo bound è inversamente proporzionale a log(1 + SNR): raddoppiare il SNR è equivalente, al primo ordine per SNR >> 1, a raddoppiare m. Più precisamente, se STP aumenta il SNR da SNR_0 a $\varrho_SNR \cdot SNR_0$ per qualche fattore $\varrho_SNR > 1$, il numero di campioni necessari si riduce secondo:

$$m_STP / m_0 \approx \log(1 + SNR_0) / \log(1 + \varrho_SNR \cdot SNR_0).$$

Per $\varrho_STP \approx 16$ si ha, invertendo questa relazione, ϱ_SNR tale che $\log(1 + \varrho_SNR \cdot SNR_0) \approx 16 \cdot \log(1 + SNR_0) / \log(1 + SNR_0) = 16$, ovvero $1 + \varrho_SNR \cdot SNR_0 \approx e^{16 \cdot \log(1 + SNR_0)}$ — un incremento sostanziale del SNR effettivo. Il Corollario H.3 della stessa appendice stabilisce il correlato sull'accuratezza: la probabilità di errore soddisfa

$$P_e \gtrsim (H(Y) - m \cdot (1/2) \log(1 + SNR)) / \log |V|,$$

dove |V| è la dimensione del vocabolario. A parità di m, un SNR più alto implica un P_e più basso (accuratezza superiore); a parità di P_e obiettivo, un SNR più

alto implica un m più piccolo (efficienza dei dati superiore). L'evidenza sperimentale valida entrambi gli effetti: STP migliora l'accuratezza rispetto al fine-tuning standard con lo stesso numero di dati (sezione 4.2 del paper STP), e mantiene l'accuratezza con 1/16 dei dati (sezione 4.3).

Come si traducono questi risultati nelle stime di compute per AMI? La sezione VI.2 aveva stimato il compute necessario per il Perception Module (I-JEPA a scala ViT con $N_{perception} \approx 10^9$ parametri su $D_{perception} \approx 10^{12}$ token visivi) come:

$C_{perception} \approx 6 \cdot N_{perception} \cdot D_{perception} \approx 6 \cdot 10^9 \cdot 10^{12} = 6 \times 10^{21}$ FLOPs.

Se l'obiettivo di training del Perception Module incorpora $\mathcal{L}_{STP}$ come regolarizzatore geometrico delle traiettorie latenti, il numero effettivo di token necessari si riduce a $D_{perception} / \varrho_{STP} \approx 10^{12} / 16 \approx 6.25 \times 10^{10}$ token, portando a:

$C_{perception}^{STP} \approx 6 \cdot 10^9 \cdot 6.25 \times 10^{10} \approx 3.75 \times 10^{20}$ FLOPs,

una riduzione di un fattore 16 rispetto alla stima originaria, corrispondente a circa 4×10^{20} FLOPs — meno di mezzo ZetaFLOP, non lontano dal compute dei run di addestramento più grandi già realizzati alla data di scrittura (GPT-4 è stimato tra 10^{23} e 10^{24} FLOPs a seconda delle stime, ma run di fine-tuning di modelli di visione di scala minore si trovano tra 10^{20} e 10^{21}). Questa riduzione non cambia l'ordine di grandezza della stima ma ne riduce significativamente il prefisso, spostando il Perception Module da "irraggiungibile con le risorse del prossimo decennio" a "al limite superiore delle capacità correnti in un orizzonte di cinque anni".

Analogamente, la stima del World Model:

$C_{worldmodel} \approx 6 \cdot 10^{10} \cdot 10^9 \cdot T_{context} \approx 6 \times 10^{23}$ FLOPs

presuppone che il modello debba essere addestrato su 10^9 traiettorie di interazione di lunghezza $T_{context}$. Se il World Model incorpora vincoli geodetici sulle traiettorie latenti — un'estensione naturale di $\mathcal{L}_{STP}$ al contesto sequenziale del World Model, dove le traiettorie sono percorsi nello spazio degli stati $\hat{s}_t$ piuttosto che sequenze di token di testo — allora anche qui il numero di traiettorie necessarie potrebbe ridursi per un fattore comparabile a ϱ_{STP}, portando a:

$C_{worldmodel}^{STP} \approx 6 \times 10^{23} / 16 \approx 3.75 \times 10^{22}$ FLOPs.

Questa stima va trattata con cautela: il risultato empirico di STP è stato ottenuto su task di fine-tuning linguistico (NL-RX-SYNTH, GSM8K, Spider, NQ-Open, HellaSwag), non su apprendimento di dinamiche di World Model. L'estensione del fattore $\varrho_{STP} \approx 16$ a questo regime richiede verifica sperimentale diretta che — alla data di completamento di questa monografia — non è ancora disponibile in letteratura. È ragionevole supporre che un qualche miglioramento di efficienza

sussista (il meccanismo di miglioramento del SNR è indipendente dal tipo di sequenza), ma il valore esatto potrebbe differire.

VI.10 — LeJEPA e la riduzione del costo di selezione del modello

Le stime di compute delle sezioni VI.1–VI.2 includono implicitamente il costo della selezione degli iperparametri: per ogni configurazione di addestramento (learning rate, weight decay, numero di epoche, coefficienti di regolarizzazione), deve essere eseguito almeno un run di validazione per valutare la qualità delle rappresentazioni apprese, tipicamente attraverso probing supervisionato su un dataset etichettato. Questo costo è raramente quantificato esplicitamente nelle stime di compute dei paper, ma in pratica rappresenta una frazione significativa del costo totale: per modelli di grande scala, ogni run di validazione richiede un'inferenza completa sull'intero validation set, e la ricerca degli iperparametri richiede decine o centinaia di run. Per sistemi della complessità di AMI, questo overhead di selezione potrebbe facilmente raddoppiare il compute totale rispetto al singolo run ottimale.

LeJEPA elimina sostanzialmente questo overhead attraverso una proprietà unica rispetto ai sistemi JEPA precedenti: la sua loss di training è altamente informativa della performance downstream, con correlazione di Spearman tra 85% e 99% a seconda del dataset e dell'architettura (sezione 6.2 del paper LeJEPA). Questa correlazione è stabile attraverso iperparametri diversi (learning rate, weight decay, numero di epoche, λ) e architetture diverse (ViT, ResNet, ConvNeXt), il che significa che una semplice ordinamento dei run secondo la loss di training fornisce, con alta affidabilità, un ordinamento corretto secondo la performance downstream. La perdita di training sostituisce il probing supervisionato come segnale di selezione del modello.

Il meccanismo teorico di questa correlazione emerge dalla struttura di LeJEPA: la loss complessiva è

$\mathcal{L}_\text{LeJEPA}(\{x_\{n,v\}\}\{n,v\}) = (\lambda/V)\ \Sigma\{v=1\}\^V\ \text{SIGReg}(\{z_\{n,v\}\}\{n=1\}\^B) + ((1-\lambda)/B)\ \Sigma\{n=1\}\^B\ \mathcal{L}_\text{pred}\^\{V_g\}(\{z_\{n,v\}\}_\{v=1\}\^V)$,

dove il termine di predizione $\mathcal{L}_\text{pred}$ misura la distanza media degli embedding di tutte le viste dal loro centroide:

$\mathcal{L}_\text{pred}(\{z_\{n,v\}\}\{v=1\}\^V) = (1/V)\ \Sigma\{v=1\}\^V\ \|\mu_n - z_\{n,v\}\|^2_2$, con $\mu_n = (1/V_g)\ \Sigma_\{v=1\}\^\{V_g\}\ z_\{n,v\}$.

Una derivazione algebrica mostra l'equivalenza tra questa formulazione e la somma normalizzata delle distanze a coppie delle prime V_g viste:

$(1/V)\ \Sigma_\{v'=1\}\^V\ \|\mu_n - z_\{n,v'\}\|^2_2 = (1/(V_g\ V))\ \Sigma_\{v=1\}\^\{V_g\}\ \Sigma_\{v'=1\}\^V\ \|z_\{n,v\} - z_\{n,v'\}\|^2_2$,

la cui dimostrazione parte dall'espansione $\|\mu_n - z_\{n,v'\}\|^2_2 = \|\mu_n\|^2_2 - 2\mu_n\^T\ z_\{n,v'\} + \|z_\{n,v'\}\|^2_2$, si sostituisce $\|\mu_n\|^2_2 = (1/V_g^2)\ \|\Sigma_v\ z_\{n,v\}\|^2_2 =$

$(1/V_g)\ \Sigma_v \|z_{n,v}\|^2_2$ (dove l'ultima uguaglianza usa la bilinearità del prodotto interno e il fatto che i termini di cross-prodotto sommano a $(1/V_g^2)\ \Sigma_{v,v''} z_{n,v}^T z_{n,v''} = \|\mu_n\|^2_2$), ottenendo la relazione dichiarata dopo sostituzione e semplificazione. Il termine SIGReg è minimizzato quando la distribuzione degli embedding è gaussiana isotrópica, che per il Teorema 1 di LeJEPA è l'ottimizzatore del rischio downstream. Il punto di minimo congiunto della loss è dunque il punto in cui le rappresentazioni sono simultaneamente isotropiche (minimizzato SIGReg) e predittive attraverso le viste (minimizzato $\mathcal{L}$_pred), che corrisponde esattamente alla condizione di massima qualità per i compiti downstream. La correlazione tra loss di training e performance downstream emerge naturalmente da questa struttura: scendere lungo la loss di training significa avvicinarsi alla configurazione ottima per i compiti downstream.

La quantificazione del risparmio computazionale dipende dal numero di run di validazione tipicamente necessari per la selezione degli iperparametri. Per il Perception Module di AMI a scala ViT-L con addestramento su ImageNet-1K, un processo di ricerca degli iperparametri standard richiederebbe tipicamente 10–30 run esplorativi a scala ridotta più 3–5 run a piena scala, per un costo aggiuntivo stimabile tra 20% e 50% del run ottimale. Con LeJEPA, poiché la loss di training è sufficiente per l'ordinamento, il numero di run a piena scala si riduce a 1–2 (il migliore dei run esplorativi scalato alla dimensione completa), con un risparmio di circa 3–4x sui run a piena scala. Per il sistema AMI completo — dove la selezione degli iperparametri deve essere effettuata per ogni modulo separatamente — questo risparmio si moltiplica per il numero di moduli, potenzialmente riducendo il costo totale di selezione del modello dal 30–50% al 5–10% del compute totale.

VI.11 — Il pretraining in-domain come alternativa al transfer learning di frontier models: implicazioni per il compute di AMI specializzato

Il risultato di LeJEPA sull'in-domain pretraining (sezione 6.3 del paper) modifica una delle assunzioni implicite più radicate nella stima del compute per sistemi AMI specializzati. L'assunzione era che la componente percettiva di un sistema AMI dovesse necessariamente essere inizializzata da un foundation model general-purpose (ViT-g addestrato su JFT-3B o simile, con compute dell'ordine di 10^{23}–10^{24} FLOPs) e poi adattata al dominio target attraverso fine-tuning. Questa strategia sembra naturale date le dimensioni del gap tra i modelli general-purpose e qualsiasi dataset in-domain, ma ha due limitazioni strutturali: il costo di utilizzo dei foundation models general-purpose è proibitivo per gruppi di ricerca con risorse limitate, e i modelli addestrati su immagini naturali non sempre generalizzano efficacemente a domini visivamente distanti come immagini astronomiche, biomedicali, o spettroscopia industriale.

LeJEPA sfida questa assunzione dimostrando che, su Galaxy10 (11.000 immagini astronomiche di 10 tipi di galassie), un ResNet-34 o ConvNeXt-V2 Nano addestrato da zero con LeJEPA per 400 epoche supera DINOv2 ViT-S (addestrato su LVD-142M, un dataset di 142 milioni di immagini naturali) sia in linear probing sia in full fine-tuning, attraverso tutto il regime di few-shot (da 1

campione a tutto il training set). I numeri specifici: LeJEPA ResNet-34 con full fine-tuning raggiunge 83,28% su Galaxy10 contro 78,34% di DINOv2 ViT-S — un gap di circa 5 punti percentuali a favore del modello in-domain molto più piccolo. Questo risultato vale anche su flowers102 (1.020 immagini), dove LeJEPA con poche centinaia di epoche produce rappresentazioni competitive rispetto ai frontier models addestrati su miliardi di immagini.

Le implicazioni per le stime di compute di AMI sono rilevanti lungo due dimensioni. La prima riguarda il compute assoluto: se il Perception Module di un sistema AMI specializzato per un dominio specifico (robotica industriale, diagnostica medica, esplorazione spaziale) può essere addestrato in-domain su dataset di 10^4–10^5 immagini anziché richiedere il transfer da un foundation model addestrato su 10^8–10^9 immagini generiche, il compute del modulo percettivo si riduce drasticamente. Per un ResNet-34 con 21M parametri addestrato su 11.000 immagini per 400 epoche con batch size 256:

$C_{perception_indomain} \approx 6 \cdot 2.1 \times 10^7 \cdot 11.000 \cdot 400 \cdot (11.000/256) \approx 6 \cdot 2.1 \times 10^7 \cdot 1.88 \times 10^7 \approx 2.37 \times 10^{15}$ FLOPs,

ovvero dell'ordine di 2 PetaFLOPs — sei ordini di grandezza inferiore alla stima di $C_{perception} \approx 6 \times 10^{21}$ FLOPs per il Perception Module general-purpose. Il divario è immenso, ma va interpretato nel contesto: il Perception Module in-domain non ha le capacità generali del fondation model, e la sua applicabilità è limitata al dominio specifico. Il punto cruciale non è che LeJEPA rende il Perception Module general-purpose economico — non lo fa — ma che per applicazioni AMI dove il dominio è noto e stabile, non è necessario investire nel Perception Module general-purpose.

La seconda dimensione riguarda il tempo di sviluppo. I ricercatori che lavorano su sistemi AMI specializzati possono iterare molto più velocemente se il Perception Module può essere addestrato in-domain in poche ore su hardware accessibile (un singolo V100 o A100) piuttosto che richiedere mesi su cluster di centinaia di GPU. La riduzione del tempo di iterazione accelera il ciclo di ricerca e permette di esplorare un numero molto maggiore di configurazioni architetturali prima di convergere sulla configurazione finale. LeJEPA rende questo tipo di esplorazione praticabile per gruppi di ricerca universitari o laboratori aziendali di piccola dimensione, democratizzando l'accesso all'architettura AMI al di là dei grandi laboratori che possono permettersi i compute dei foundation models general-purpose.

VI.12 — Revisione delle stime di sostenibilità energetica

La sezione VI.5 della monografia aveva discusso la sostenibilità energetica dei sistemi AMI di grande scala, concludendo che il compute necessario per un sistema AMI a piena scala (dell'ordine di 10^{23}–10^{24} FLOPs) implica consumi energetici comparabili al PIL di nazioni di medie dimensioni e che questo

rappresenta un limite pratico non trascurabile. I risultati di LeJEPA e STP modificano queste stime in modo significativo lungo tre dimensioni.

La prima dimensione è la riduzione dell'impronta energetica del Perception Module. Con il fattore $\varrho_STP \approx 16$ nell'efficienza dei dati, il compute del Perception Module general-purpose si riduce da $C_perception \approx 6 \times 10^{21}$ FLOPs a $C_perception^{STP} \approx 3.75 \times 10^{20}$ FLOPs. Usando il fattore di conversione approssimativo di 1 exaFLOP ≈ 7 kWh (stimato da analisi empiriche del consumo di data center con A100, considerando un'efficienza di picco di $\sim 3 \times 10^{14}$ FLOP/s/W e un overhead di raffreddamento di circa 2x), si ha:

$$E_perception^{STP} \approx 3.75 \times 10^{20} / 10^{18} \cdot 7 \text{ kWh} \approx 375 \cdot 7 \approx 2.625 \text{ MWh},$$

contro $E_perception \approx 6 \times 10^{21} / 10^{18} \cdot 7 \approx 42$ MWh della stima originale. Il risparmio di circa 40 MWh è equivalente al consumo annuale di circa quattro appartamenti europei — non banale ma non ancora dell'ordine di grandezza necessario per cambiare la discussione sulla sostenibilità a livello di sistema.

La seconda dimensione è la più rilevante sul lungo periodo: la qualità dell'apprendimento per campione. La sezione VI.5 della monografia aveva argomentato che una delle vie verso una maggiore sostenibilità è lo spostamento verso paradigmi di apprendimento radicalmente più efficienti dal punto di vista dei dati, usando come base la congettura che un sistema AMI con World Model interattivo potrebbe imparare da 10^9 step di interazione anziché 10^{13} token passivi. STP offre per la prima volta una prova empirica e un meccanismo teorico preciso per questa congettura: l'aggiunta di un termine di regolarizzazione geometricamente motivato all'obiettivo di training aumenta il SNR di ogni campione, riducendo il numero di campioni necessari per la convergenza. Il meccanismo non dipende dal tipo di dati (sequenze di testo, traiettorie di interazione, sequenze video) ma dalla struttura latente dello spazio di rappresentazione — che STP ipotizza essere geodeticamente organizzato per qualsiasi dominio dove il Principio di Minima Azione vale. Se questa ipotesi si generalizza oltre i task linguistici su cui è stata validata, il fattore $\varrho_STP \approx 16$ rappresenterebbe un risparmio energetico dell'ordine di 94% rispetto ai paradigmi di training correnti per qualsiasi componente del sistema AMI che possa beneficiare di questa regolarizzazione.

La terza dimensione riguarda la riduzione dell'overhead di selezione del modello discussa in VI.10. Un'analisi della letteratura sul training di modelli di linguaggio di grande scala (Hoffmann et al., 2022; Grattafiori et al., 2024) suggerisce che la fase di ablazione e selezione degli iperparametri tipicamente consuma tra il 20% e il 40% del compute del run finale. Per un sistema AMI con $C_total \approx 6 \times 10^{23}$ FLOPs, questo overhead è dell'ordine di $1.2–2.4 \times 10^{23}$ FLOPs, corrispondenti a 840–1.680 MWh di energia. L'eliminazione della necessità di probing supervisionato per la selezione del modello, resa possibile dalla loss informativa di LeJEPA, riduce questo overhead di un fattore stimabile tra 3 e 5 — un risparmio

dell'ordine di 500–1.300 MWh che, seppure non modifichi l'ordine di grandezza della stima totale, è significativo in termini assoluti.

Combinando i tre effetti, una stima revisionata del compute totale per un sistema AMI di media scala che incorpori LeJEPA e STP come componenti standard è:

C_AMI_revised ≈ C_perception^{STP} + C_worldmodel^{STP} + C_overhead_ridotto ≈ $3.75 \times 10^{20} + 3.75 \times 10^{22}$ + (riduzione 3-5x degli overhead) ≈ $\sim 4 \times 10^{22}$ FLOPs per il training principale,

con una riduzione rispetto alla stima originale ($\approx 6 \times 10^{23}$ FLOPs) di un fattore 15, ovvero quasi due ordini di grandezza. In termini energetici, questo porta la stima da circa 4.200 MWh a circa 280 MWh — ancora un consumo molto elevato in termini assoluti, ma già all'interno del range raggiunto da training run esistenti di modelli di frontiera e non più di un ordine di grandezza superiore al consumo annuale di una piccola istituzione di ricerca.

Va sottolineato che queste stime revisionate si fondano sull'estrapolazione del fattore ϱ_STP ≈ 16 al di là del dominio sperimentale verificato (fine-tuning linguistico su task sintetici e dataset standard) verso il regime del pre-training su larga scala e del training di World Models su traiettorie di interazione. Si tratta di estrapolazioni teoricamente motivate ma non ancora validate empiricamente. Le Scaling Laws per STP — la forma funzionale di ϱ_STP come funzione di N, D e dell'architettura — non sono state ancora caratterizzate, e il comportamento di ϱ_STP nel limite di grande scala ($N > 10^{10}$, $D > 10^{12}$) rimane un'aperta incognita. La Previsione P4 di STP ($\lambda \ll 1$ ottimale) implica che per architetture più grandi e dati più abbondanti la geodetica devia di più dalla retta lineare — la curvatura del manifold semantico aumenta con la complessità del dominio — il che potrebbe ridurre ϱ_STP per modelli very large. Questa è, al momento, la principale incertezza nella revisione delle stime di sostenibilità.

VI.13 — Implicazioni per gli scenari temporali dell'AGI

Le stime revisionate del compute necessario per AMI, alla luce dei risultati di LeJEPA e STP, permettono di aggiornare i tre scenari temporali delineati in VI.7 con dati quantitativi più precisi.

Nello scenario di continuazione tendenziale, la riduzione del compute per il Perception Module da 6×10^{21} a $\sim 4 \times 10^{20}$ FLOPs significa che questo componente è già potenzialmente realizzabile con le risorse hardware disponibili al momento della scrittura (un cluster di 1.000 A100 per circa 30 giorni, a un costo dell'ordine di 10^6 USD al prezzo di mercato del cloud computing). La riduzione del costo del World Module da 6×10^{23} a $\sim 4 \times 10^{22}$ FLOPs lo porta nell'ordine di grandezza di training run effettivamente eseguiti da laboratori di frontiera (GPT-4 si colloca intorno a 10^{23} FLOPs secondo le stime), il che suggerisce che un prototipo di sistema AMI a scala ridotta potrebbe essere realizzabile nell'arco di tre-cinque anni anziché dieci-venti come stimato nello scenario originale. Questo

aggiornamento sposta lo scenario di continuazione tendenziale verso quello che in VI.7 era classificato come "scenario ottimistico".

Nello scenario ottimistico, il breakthrough algoritmico che prima si ipotizzava nella forma di un meccanismo di apprendimento radicalmente più efficiente è ora parzialmente disponibile nella forma di STP e LeJEPA, che forniscono il meccanismo teorico (miglioramento del SNR attraverso vincoli geometrici) e l'implementazione pratica ($\mathcal{L}$_STP e SIGReg come termini di regolarizzazione). La domanda critica è se il fattore ϱ_STP $\approx$ 16 sia un'anomalia specifica ai task linguistici su cui è stato misurato o un effetto fondamentale che scala a qualsiasi dominio dove la struttura latente è geodeticamente organizzata. La risposta — che non è ancora disponibile alla data di stesura di questa monografia — determinerà se lo scenario ottimistico è ora lo scenario atteso o se rimane nello stesso orizzonte di incertezza che aveva prima di questi risultati.

Nello scenario pessimistico, le nuove evidenze non riducono significativamente l'incertezza: se esistono barriere di principio alla composizionalità sistematica o al symbol grounding (discusse in V.8) che non sono risolte da miglioramenti nell'efficienza dei dati, allora il fattore ϱ_STP è irrilevante per l'orizzonte AGI nel senso pieno. STP e LeJEPA migliorano l'efficienza dell'apprendimento per campione, ma se il problema fondamentale non è la quantità di dati bensì il tipo di struttura cognitiva richiesta — composizionalità algebrica, ragionamento causale esplicito, grounding nell'azione embodied — allora quindici volte meno dati con la stessa architettura non cambiano la situazione qualitativa.

La posizione epistemicamente più onesta è che i risultati di LeJEPA e STP aggiornano in modo significativo le stime quantitative del compute necessario (verso il basso di un fattore ~15), non modificano sostanzialmente la struttura qualitativa degli scenari (le barriere fondamentali rimangono quelle identificate in V.8), e accelerano di alcuni anni l'orizzonte temporale degli scenari ottimistico e di continuazione tendenziale riducendo la barriera del compute da "irraggiungibile oggi" a "raggiungibile in orizzonte di tre-cinque anni per prototipi a scala ridotta". Lo scenario pessimistico rimane invariato perché la sua natura è di principio, non di costo. Questa distinzione — tra progress sul compute efficiency e progress sulle barriere fondamentali — è la misura del contributo effettivo che i risultati presentati in questa sezione integrativa apportano alla valutazione complessiva del progetto AMI.

Capitolo VII — JEPA come Tokenizzatore Neurale per il Parlato: Attenzione Adattiva di Densità e Quantizzazione Scalare Finita in un Regime Autosupervisionato a Due Stadi

VII.1 — Il problema della rappresentazione sonora come caso particolare del problema generale della rappresentazione

La monografia ha fino a questo punto sviluppato il problema della rappresentazione principalmente nel dominio visivo, dove il paradigma JEPA è stato formalizzato nella sua variante originale e nei suoi successori teorici — LeJEPA con la regolarizzazione gaussiana isotrópica, V-JEPA con l'estensione spaziotemporale al video, MC-JEPA con la struttura multi-scala. La domanda se i principi che motivano JEPA — separazione fra apprendimento delle rappresentazioni e ricostruzione sensoriale, predizione nello spazio latente, prevenzione del collasso mediante target encoder aggiornato con media esponenziale mobile — si trasferiscano a domini modali diversi non è, nella prospettiva del framework AMI, una domanda accessoria. Un sistema di intelligenza autonoma che aspiri a operare nel mondo reale deve integrare segnali provenienti da molteplici canali sensoriali, e l'architettura del Perception Module non può permettersi di avere componenti teoricamente ben fondati per la visione ma empiricamente ad hoc per l'udito, il tatto o la propriocezione.

Il segnale vocale presenta rispetto all'immagine statica differenze strutturali profonde che impongono rielaborazioni non banali del paradigma JEPA. La prima differenza è la natura intrinsecamente temporale del segnale: mentre un'immagine è un campo scalare o vettoriale su un dominio bidimensionale discreto e stazionario, una forma d'onda audio è una sequenza monovariata campionata tipicamente a 22 050 Hz o 24 000 Hz, il che implica che un secondo di parlato sia rappresentato da un vettore di circa 24 000 dimensioni prima di qualsiasi elaborazione. La seconda differenza è la struttura gerarchica multiscala delle caratteristiche rilevanti: il parlato contiene simultaneamente informazione fonetica a scale temporali dell'ordine dei 10–50 millisecondi, informazione prosodica e sillabica a scale dell'ordine dei 100–500 millisecondi, e informazione semantica e discorsiva a scale di secondi. Nessuna finestra di analisi di lunghezza fissa cattura simultaneamente tutte queste scale, il che rende la progettazione della strategia di mascheramento e dell'architettura dell'encoder compiti qualitativamente più delicati che nel dominio visivo. La terza differenza è la multimodalità della distribuzione del segnale condizionalmente al testo: dato un enunciato semanticamente specificato, esistono innumerevoli realizzazioni acustiche plausibili che differiscono in timbro, altezza, velocità, accento, emozione espressiva — varianti tutte semanticamente equivalenti ma fisicamente distinte che un buon encoder dovrebbe collapsa in rappresentazioni latenti simili.

Il lavoro di Ioannides, Constantinou, Chadha, Elkins, Pang, Shwartz-Ziv e LeCun, pubblicato nell'autunno del 2025 con il titolo "JEPA as a Neural Tokenizer: Learning Robust Speech Representations with Density Adaptive Attention", affronta queste sfide attraverso un'architettura a due stadi che estende il principio JEPA al dominio del parlato integrando un meccanismo di attenzione adattiva basato su misture gaussiane — denominato Density Adaptive Attention Mechanism (DAAM) — e un sistema di quantizzazione e ricostruzione che converte le rappresentazioni continue apprese nel primo stadio in token discreti adatti all'ingestione da parte di modelli linguistici. Il risultato è un sistema che produce 47,5 token al secondo — un tasso di compressione inferiore di uno o due ordini di grandezza rispetto alla maggior parte dei codec audio neurali esistenti — mantenendo al contempo la reversibilità dell'operazione di tokenizzazione e la compatibilità con i vocabolari di dimensione comparabile a quelli adoperati nell'elaborazione del linguaggio naturale.

VII.2 — Motivazione teorica: perché JEPA per il parlato e non un autoencoder variazionale

La tradizione del codec audio neurale — che va da SoundStream di Zeghidour et al. (2021) a EnCodec di Défossez et al. (2022), da DAC di Kumar et al. (2024) a Mimi — ha adottato sistematicamente un paradigma che accoppia l'apprendimento delle rappresentazioni con la ricostruzione della forma d'onda. In questi sistemi, l'encoder apprende a comprimere il segnale audio in un codice discreto e il decoder apprende a ricostruirlo con fedeltà percettiva più alta possibile, e i due moduli vengono ottimizzati congiuntamente rispetto a perdite che misurano la qualità della ricostruzione nello spazio delle osservazioni — tipicamente una combinazione di perdita L1 sulla forma d'onda, perdita spettrale multi-risoluzione e perdita avversariale da discriminatori GAN.

Il problema di questa impostazione, che nel Capitolo I era stato analizzato per il caso visivo, si ripresenta nel caso audio con ulteriore acuità. Minimizzare la perdita di ricostruzione della forma d'onda costringe l'encoder a preservare informazioni che sono irrilevanti per il contenuto semantico del segnale ma necessarie per la fedeltà acustica: le fluttuazioni di fase delle componenti ad alta frequenza, il jitter periodico del pitch, le caratteristiche del rumore di fondo, la firma spettrale del microfono e dell'ambiente di registrazione. Un encoder ottimizzato rispetto alla fedeltà di ricostruzione dedica quindi capacità rappresentativa a caratteristiche acustiche che non hanno valore per la comprensione del parlato, a scapito delle strutture linguistiche e prosodiche che portano il contenuto informativo rilevante.

JEPA risolve questo problema attraverso la stessa separazione architettonica che aveva motivato I-JEPA nel dominio visivo: il Primo Stadio apprende rappresentazioni latenti esclusivamente attraverso la predizione nello spazio latente, con una perdita MSE tra predizioni e target definiti sulle rappresentazioni di un target encoder aggiornato via EMA, senza nessun termine di ricostruzione della forma d'onda. Soltanto nel Secondo Stadio le rappresentazioni apprese nel

Primo vengono utilizzate come punto di partenza per addestrare un decoder HiFi-GAN che ricostruisce la forma d'onda originale. La separazione fra i due obiettivi — semantico il primo, percettivo-acustico il secondo — è l'innovazione architettonica centrale del sistema, e la sua plausibilità teorica si fonda sugli stessi argomenti dell'Information Bottleneck sviluppati nel Capitolo II per il caso visivo: l'encoder del Primo Stadio è libero di ottimizzare unicamente la qualità predittiva delle sue rappresentazioni, senza dover rispondere alle esigenze spesso conflittuali della fedeltà di ricostruzione.

La presenza del meccanismo DAAM nell'encoder del Primo Stadio introduce un ulteriore grado di libertà nell'ottimizzazione: l'encoder non soltanto apprende a predire le rappresentazioni delle regioni mascherate, ma impara simultaneamente a identificare le regioni temporali più statisticamente salienti attraverso un sistema di gate calcolato a partire da misture di gaussiane sui dati. Questo meccanismo, originariamente proposto da Ioannides, Chadha e Elkins nel 2024 nel contesto del fine-tuning efficiente nei parametri attraverso modalità multiple, viene qui adattato e stabilizzato per l'integrazione nell'encoder JEPA, producendo una forma di attenzione temporale adattiva che opera su statistiche distribuzionali piuttosto che su prodotti scalari punto-a-punto come l'attenzione softmax standard.

VII.3 — Primo Stadio: il framework JEPA per sequenze audio

Il Primo Stadio del sistema è un framework JEPA autosupervisionato applicato a sequenze audio campionate a 24 000 Hz. Sia $\mathbf{x} \in \mathbb{R}^{(1 \times T_wav)}$ una forma d'onda monoaurale di lunghezza T_wav campioni. La struttura architetturale replica il pattern fondamentale di I-JEPA con le necessarie modifiche per il dominio temporale unidimensionale.

VII.3.1 — La strategia di mascheramento a blocchi temporali

Il framework JEPA per il parlato adotta una strategia di mascheramento a blocchi contigui (block masking) sulle rappresentazioni temporali a bassa frequenza di fotogramma. Sia T la lunghezza della sequenza dopo il downsampling effettuato dall'encoder convoluzionale. Le maschere binarie $\mathbf{m} \in \{0, 1\}^{(B \times T)}$ sono generate secondo il seguente algoritmo, parametrizzato da un rapporto di mascheramento $\varrho \in [0, 1]$, una lunghezza minima di span s_min e una lunghezza massima s_max.

Per ciascun campione b nel batch si inizializza $\mathbf{m}[b, :] = 1$ (tutte le posizioni visibili) e si pone n_masked $\leftarrow 0$. Si itera la seguente procedura finché n_masked $< \lfloor \varrho \cdot T \rfloor$: si campiona una lunghezza di span $\ell \sim$ Uniform(s_min, s_max), si campiona una posizione di inizio t_start $\sim$ Uniform(0, $T - \ell$), si computa la posizione di fine t_end $\leftarrow$ min(t_start + ℓ, T), si pone $\mathbf{m}[b, t] \leftarrow 0$ per ogni $t \in$ [t_start, t_end) e si aggiorna n_masked $\leftarrow$ n_masked + (t_end − t_start).

Gli iperparametri adottati nel sistema sono: $\varrho = 0{,}5$ (il cinquanta per cento dei fotogrammi è mascherato), s_min = 2 fotogrammi, s_max = T/4 fotogrammi (adattivo alla lunghezza della sequenza). A un tasso di fotogrammi di 2,5 Hz —

ottenuto attraverso un downsampling complessivo di 9 600 campioni per hop a 24 000 Hz — gli span corrispondono a intervalli temporali variabili calibrati sulla lunghezza effettiva della sequenza.

La motivazione per il mascheramento a blocchi contigui anziché a posizioni casuali indipendenti è analoga a quella sviluppata nel Capitolo III per I-JEPA nel dominio visivo: il mascheramento di regioni estese e contigue costringe il modello ad apprendere dipendenze temporali a lungo raggio e strutture semantiche globali, poiché non è possibile predire la rappresentazione di una regione mascherata attraverso interpolazione locale dalle patch o dai fotogrammi adiacenti. Nel dominio del parlato questa motivazione è particolarmente forte: una parola mascherata non può essere predetta dalla struttura spettrale locale delle sillabe immediatamente precedenti senza comprensione della struttura fonetica e prosodica dell'enunciato.

VII.3.2 — Il meccanismo di Attenzione Adattiva di Densità (DAAM)

Il DAAM è il componente architetturalmente più originale del sistema e merita una trattazione matematica dettagliata. Sia $\mathbf{x} \in \mathbb{R}^{B \times C \times T}$ il tensore di feature di input (dimensioni batch, canali, tempo). Il modulo DensityAdaptiveAttention opera lungo l'asse temporale calcolando gate di attenzione adattivi basati su distribuzioni di mistura di gaussiane apprese.

Il calcolo procede attraverso sei passi distinti.

Passo 1 — Statistiche temporali. Per ciascun campione del batch e ciascun canale si calcolano la media e la varianza attraverso la dimensione temporale:

$$\mu = (1/T) \cdot \Sigma_{t=1}^{T} \mathbf{x}_{:,:,t} \in \mathbb{R}^{B \times C \times 1}$$

$$\sigma^2 = (1/T) \cdot \Sigma_{t=1}^{T} (\mathbf{x}_{:,:,t} - \mu)^2 \in \mathbb{R}^{B \times C \times 1}$$

La varianza viene vincolata inferiormente: $\text{var} \geq 10^{-6}$ per stabilità numerica.

Passo 2 — Parametri gaussiani apprendibili. Per K componenti gaussiani (K = 4 nell'implementazione), il modulo mantiene parametri scalari apprendibili: offset di media $\delta = [\delta_1, \ldots, \delta_K] \in \mathbb{R}^K$ (inizializzati a zero) e parametri di log-scala $\nu = [\nu_1, \ldots, \nu_K] \in \mathbb{R}^K$ (inizializzati a $\log(0{,}5)$). Le scale positive vengono calcolate attraverso la funzione softplus con termine $\varepsilon = 10^{-3}$ per stabilità:

$$\tilde{\sigma}_k = \text{softplus}(\nu_k) + \varepsilon = \log(1 + \exp(\nu_k)) + \varepsilon$$

Passo 3 — Deviazioni standardizzate per componente. Per ciascun componente k e ciascun fotogramma temporale t si calcola la deviazione standardizzata rispetto alla media adattata dall'offset apprendibile:

$$z_{k,t} = (\mathbf{x}_{:,:,t} - (\mu + \delta_k)) \,/\, (\sigma \cdot \tilde{\sigma}_k + \varepsilon)$$

Passo 4 — Log-densità sotto ciascuna gaussiana. La log-probabilità di densità al fotogramma t sotto il k-esimo componente gaussiano è:

$$\log p_k(\mathbf{x}t) = -(1/2) \cdot \chi^2\{k,t\} - \log \tilde{\sigma}_k - (1/2) \cdot \log(2\pi)$$

Passo 5 — Aggregazione della mistura mediante log-sum-exp. La log-densità della mistura viene calcolata tramite l'operazione numericamente stabile logsumexp:

$$\log G(\mathbf{x}_t) = \text{logsumexp}(\{\log p_1(\mathbf{x}_t), \ldots, \log p_K(\mathbf{x}_t)\}) - \log K$$

La sottrazione di log K normalizza per il numero di componenti, rendendo il gate invariante rispetto alla scelta di K.

Passo 6 — Gate di attenzione e modulazione delle feature. Il gate di attenzione finale è ottenuto esponenzializzando la log-densità della mistura:

$$G(\mathbf{x}_t) = \exp(\log G(\mathbf{x}_t))$$

e le feature di uscita vengono calcolate attraverso modulazione elemento per elemento (prodotto di Hadamard):

$$\mathbf{y}_t = \mathbf{x}_t \odot G(\mathbf{x}_t)$$

L'implementazione pratica di DAAM opera su una proiezione monocanale: le feature vengono prima proiettate a un singolo canale attraverso una convoluzione 1×1, il gate gaussiano viene calcolato su questo segnale temporale unidimensionale, e il gate risultante viene applicato per scalare il tensore di feature completo. Nella versione integrata nell'encoder JEPA, il gate non viene applicato direttamente ma attraverso una forma residuale che controlla la forza della modulazione tramite un parametro α apprendibile inizializzato a 0,05:

$$\mathbf{y} = \mathbf{x} \cdot (1 + \alpha \cdot \text{gate})$$

Questa formula introduce una dipendenza lineare dal gate con coefficiente adattivo, permettendo al sistema di regolare autonomamente quanto la modulazione basata sulla densità contribuisce alla trasformazione delle feature rispetto all'input non modulato.

La differenza concettuale fra DAAM e l'attenzione softmax standard è profonda e teoricamente motivata. L'attenzione softmax risponde alla domanda "quali fotogrammi temporali sono più simili a questo?", calcolando correlazioni punto-a-punto tra query e chiavi proiettate. DAAM risponde invece alla domanda "quali fotogrammi temporali hanno proprietà statistiche inusuali o particolarmente informative?", identificando le regioni in cui il segnale si discosta dalla sua distribuzione temporale attesa secondo un modello di mistura gaussiana appreso. Le regioni acusticamente salienti — transizioni fonemiche, attacchi consonantici, variazioni di intonazione — tenderanno a coincidere con le posizioni dove il segnale ha alta densità sotto la mistura, producendo gate elevati che amplificano selettivamente le feature corrispondenti.

Rispetto all'attenzione softmax, DAAM non richiede il calcolo di matrici di attenzione di dimensione T × T, evitando la complessità quadratica tipica

dell'attenzione standard e rendendolo computazionalmente vantaggioso per sequenze audio lunghe. Il calcolo delle statistiche temporali e dei gate ha complessità O(B · C · T · K), lineare nella lunghezza della sequenza per K fissato.

Tutti i calcoli del DAAM vengono eseguiti in precisione FP32 per stabilità numerica, indipendentemente dalla precisione usata nel resto del forward pass.

VII.3.3 — L'architettura dell'encoder JEPA con DAAM

L'encoder JEPA è un sistema ibrido convolutivo-Transformer che processa la forma d'onda grezza e produce rappresentazioni latenti a bassa frequenza di fotogramma. L'architettura comprende due percorsi paralleli che condividono i pesi ma svolgono ruoli distinti nell'addestramento.

L'**encoder online** (rete contestuale) processa l'input audio completo. Il mascheramento viene applicato successivamente nello spazio delle feature sostituendo i fotogrammi mascherati con un token di maschera apprendibile, prima dell'ingresso nella rete predittiva. I suoi parametri vengono aggiornati attraverso discesa del gradiente rispetto alla perdita di predizione.

L'**encoder target** (rete EMA) processa anch'esso l'intera forma d'onda audio — incluse le regioni corrispondenti alle posizioni mascherate — e produce le rappresentazioni target per la supervisione. I suoi parametri vengono aggiornati esclusivamente attraverso media esponenziale mobile dall'encoder online, senza ricevere gradienti attraverso la backpropagation.

Il percorso di downsampling trasforma la forma d'onda grezza $\mathbf{x} \in \mathbb{R}^{(B \times 1 \times T_wav)}$ attraverso una sequenza di blocchi convoluzionali strided che ampliano progressivamente il numero di canali:

$64 \rightarrow 128 \rightarrow 256 \rightarrow 384 \rightarrow 512 \rightarrow 512$

con strides complessivi pari a $8 \times 8 \times 5 \times 5 \times 6 = 9\,600$ campioni per hop a 24 000 Hz, producendo una rappresentazione latente $\mathbf{z} \in \mathbb{R}^{(B \times 512 \times T_z)}$ dove T_z corrisponde a un tasso di fotogrammi di 2,5 Hz. Ciascun blocco encoder comprende una convoluzione 1D strided, l'attivazione SnakeBeta (che fornisce bias induttivo periodico appropriato per segnali audio), blocchi residuali e moduli GAttnGate con gate DAAM.

La rappresentazione convoluzionale viene quindi processata da 8 blocchi Conformer con 16 teste di attenzione ciascuno. Ogni blocco Conformer comprende attenzione multi-testa, feed-forward network, convoluzione depthwise e normalizzazione per layer — l'architettura standard dei Conformer di Gulati et al. (2020), riconosciuta come particolarmente adatta per sequenze audio per via dell'integrazione di meccanismi di attenzione globale e convoluzione locale che cattura strutture ritmiche e fonetiche a scala ridotta. I moduli GAttnGate con gate DAAM vengono applicati dopo ciascun blocco encoder strided, non dopo i blocchi Conformer nell'implementazione corrente.

VII.3.4 — La rete predittiva JEPA per il parlato

La rete predittiva (predictor) riceve le rappresentazioni del contesto prodotte dall'encoder online e deve predire le rappresentazioni delle regioni mascherate. La sua architettura comprende: uno strato di espansione Conv1D che raddoppia la dimensione dei canali, due blocchi Conformer separati da una Conv1D intermedia per il raffinamento delle feature, e uno strato di proiezione Conv1D finale che riduce alla dimensionalità originale.

Il predictor riceve esclusivamente le feature del contesto (regioni visibili) ma deve produrre predizioni per tutte le posizioni temporali, incluse quelle mascherate. La maschera viene applicata alla perdita piuttosto che all'input: il predictor non "vede" le posizioni target ma deve inferirne le rappresentazioni dal contesto latente e dagli embedding posizionali delle posizioni mascherate.

La dimensione ridotta del predictor rispetto all'encoder — il predictor utilizza due blocchi Conformer contro gli otto dell'encoder — è una scelta deliberata che riflette il principio architetturale discusso nel Capitolo III a proposito di I-JEPA: un predictor eccessivamente espressivo potrebbe apprendere trasformazioni arbitrariamente complesse delle feature del contesto, riducendo la pressione sull'encoder affinché le sue rappresentazioni siano intrinsecamente predittive. Un predictor di capacità limitata costringe implicitamente l'encoder a produrre rappresentazioni già allineate con il target nello stesso spazio semantico.

VII.3.5 — L'obiettivo di addestramento del Primo Stadio

La perdita di addestramento del Primo Stadio è una perdita MSE (errore quadratico medio) definita esclusivamente sulle posizioni mascherate nello spazio latente:

$$\mathcal{L}_{JEPA} = (1 / (N_{mask} \cdot C)) \cdot \Sigma_{t \in \mathcal{M}} \| \mathbf{z}^{(t)}_{pred} - sg(\mathbf{z}^{(t)}_{target}) \|^2$$

dove $\mathcal{M} = \{t : m_t = 0\}$ è l'insieme delle posizioni mascherate, $N_{mask} = |\mathcal{M}|$, C è la dimensione dei canali, $\mathbf{z}^{(t)}_{pred}$ è la predizione del predictor per la posizione t, $\mathbf{z}^{(t)}_{target}$ è la rappresentazione della medesima posizione prodotta dal target encoder, e $sg(\cdot)$ è l'operazione stop-gradient che impedisce ai gradienti di fluire attraverso il target encoder. La normalizzazione per $N_{mask} \cdot C$ rende la perdita indipendente dalla lunghezza della sequenza e dalla dimensionalità latente.

L'aggiornamento del target encoder viene effettuato dopo ciascun passo di addestramento secondo la formula EMA standard:

$$\theta_{target} \leftarrow \tau \cdot \theta_{target} + (1 - \tau) \cdot \theta_{online}$$

con coefficiente di momento $\tau = 0{,}996$. Questo valore, selezionato attraverso la pratica empirica consolidata della letteratura JEPA e BYOL, mantiene il target encoder sufficientemente stabile da produrre target coerenti senza essere così statico da rendere il compito predittivo trivialmente semplice.

Gli iperparametri di addestramento del Primo Stadio sono: ottimizzatore AdamW con $\beta_1 = 0{,}8$, $\beta_2 = 0{,}99$; tasso di apprendimento $1{,}5 \times 10^{-4}$; weight decay 10^{-3}; batch size 32; lunghezza massima audio 15 secondi a 24 kHz; 24 000 passi di addestramento totali.

Il sistema implementa il monitoraggio del collasso rappresentazionale misurando, senza backpropagation, la deviazione standard delle uscite del predictor attraverso le dimensioni batch e temporali. Qualora la deviazione standard media scenda al di sotto di 0,01, viene registrato un avviso diagnostico. Questo meccanismo non contribuisce alla perdita ma fornisce un segnale di allarme precoce che indica l'eventuale degenerazione delle rappresentazioni verso la soluzione collassata.

VII.4 — Secondo Stadio: quantizzazione scalare finita, impacchettamento misto-radix e ricostruzione HiFi-GAN

Al termine del Primo Stadio, i pesi dell'encoder JEPA vengono utilizzati come punto di partenza per il Secondo Stadio, nel quale vengono fine-tuned congiuntamente con un quantizzatore FSQ e un decoder HiFi-GAN rispetto a obiettivi di ricostruzione della forma d'onda.

VII.4.1 — La quantizzazione scalare finita (FSQ)

La quantizzazione scalare finita (Finite Scalar Quantization, FSQ), introdotta da Mentzer et al. nel 2023, è un metodo di tokenizzazione discreta che evita l'apprendimento di codebook espliciti tipico del VQ-VAE. In luogo di un codebook apprendibile, FSQ adopera quantizzazione scalare fissa per dimensione, il che elimina i problemi di instabilità nell'aggiornamento del codebook e di bassa utilizzazione che affliggono i metodi VQ in alta dimensione.

Sia $\mathbf{z}_e \in \mathbb{R}^{B \times C \times T}$ il tensore di feature dell'encoder. La procedura di quantizzazione FSQ procede in due passi.

Proiezione. Le feature vengono prima proiettate nell'intervallo $[-1, 1]$ mediante la funzione tangente iperbolica:

$$\mathbf{z}'_e = \tanh(\mathbf{z}_e)$$

Quantizzazione per dimensione. Per la dimensione d con livello di quantizzazione L_d, si definisce l'insieme dei valori di confine:

$$B_d = \{ (2i - L_d + 1) / L_d : i \in \{0, 1, \ldots, L_d - 1\} \}$$

La funzione di quantizzazione assegna a ciascun valore scalare il confine più vicino:

$$q_d(x) = \operatorname{argmin}_{b \in B_d} |x - b|$$

e il valore quantizzato è $\mathbf{z}_q[d] = q_d(\mathbf{z}'_e[d])$.

La configurazione adottata nel sistema è L = [4, 4, 4, 4] (quattro livelli per ognuna di quattro dimensioni rilevanti), dimensione del codice C = 128 e temperatura τ = 1,0. Con quattro livelli di quantizzazione per dimensione e 128 dimensioni, lo spazio di codifica ha in principio 4^{128} possibili configurazioni distinte — un numero astronomico che viene reso gestibile attraverso la strategia di impacchettamento misto-radix descritta nella sezione seguente.

La propagazione del gradiente attraverso l'operazione di quantizzazione — che è a gradiente nullo quasi ovunque — viene gestita mediante lo stimatore straight-through: durante la backpropagation il gradiente viene propagato attraverso la quantizzazione come se questa fosse la funzione identità:

$\partial \mathcal{L} / \partial \mathbf{z}_e = \partial \mathcal{L} / \partial \mathbf{z}_q$

Questa approssimazione, introdotta da Bengio et al. nel contesto degli autoencoders binari e ampiamente adottata nella letteratura dei codec neurali, permette di addestrare l'encoder e il decoder congiuntamente nonostante la non-differenziabilità dell'operazione di quantizzazione.

VII.4.2 — L'impacchettamento misto-radix dei token FSQ

L'algoritmo di impacchettamento misto-radix (mixed-radix token packing) è il componente che converte gli indici FSQ in token interi compatti adatti all'elaborazione da parte di modelli linguistici. L'intuizione fondamentale è che un insieme di indici in una base mista può essere codificato come un singolo intero mediante aritmetica modulare, esattamente come un sistema posizionale numerico — solo che in questo caso la base può variare da posizione a posizione.

Sia $\mathbf{i} \in \mathbb{Z}^{(B \times T \times D)}$ il tensore degli indici FSQ, con raggi (radices) specifici per dimensione $\mathbf{r} = [r_1, \ldots, r_G]$ per un gruppo di G dimensioni. Qualsiasi combinazione $[i_1, \ldots, i_G]$ viene codificata come:

$$\text{token} = \Sigma_{k=1}^{G} i_k \cdot \Pi_{j=k+1}^{G} r_j$$

Questa formula è la rappresentazione posizionale misto-radix: la k-esima componente contribuisce al token con peso pari al prodotto di tutti i raggi delle posizioni successive. A titolo illustrativo, per G = 7 e $\mathbf{r}$ = [4, 4, 4, 4, 4, 4, 4] con $\mathbf{i}$ = [2, 1, 3, 0, 2, 1, 3] il token vale:

$$\text{token} = 2 \cdot 4^6 + 1 \cdot 4^5 + 3 \cdot 4^4 + 0 \cdot 4^3 + 2 \cdot 4^2 + 1 \cdot 4^1 + 3 \cdot 4^0 = 10\,023$$

con valore massimo $4^7 - 1 = 16\,383$, corrispondente a un vocabolario di 16 384 elementi — dimensione comparabile ai vocabolari subword adoperati nell'elaborazione del linguaggio naturale.

Il calcolo efficiente del token viene realizzato mediante il metodo di Horner, che riduce il numero di moltiplicazioni da $O(G^2)$ a $O(G)$:

token = $i_1 \cdot r_2 \cdot r_3 \cdot \ldots \cdot r_G + \ldots + i_\{G-1\} \cdot r_G + i_G$

implementato iterativamente da destra a sinistra: si inizializza token ← i_G, poi per k = G−1 fino a 1 si esegue token ← i_k + token · r_k. Questo algoritmo richiede esattamente G−1 moltiplicazioni e G addizioni.

L'implementazione FSQ dell'intero sistema produce D = 128 dimensioni quantizzate. Con dimensione del gruppo G = 7, il numero di gruppi necessari per coprire le 128 dimensioni è $\lceil 128/7 \rceil = 19$, con $19 \times 7 - 128 = 5$ dimensioni di padding con raggio 1 (che non contribuiscono al valore del token). Il tasso di fotogrammi complessivo è:

f = (sample rate) / hop = 24 000 / 9 600 = 2,5 Hz

con 19 gruppi per fotogramma, il tasso di token è:

tps = $2{,}5 \times 19 = 47{,}5$ token al secondo

L'operazione di decodifica (conversione da token a indici) è l'inversa dell'encoding e procede per divisione intera e modulo: si inizializza rem ← token, poi per k = 1 fino a G si calcola prod = $\Pi_\{j=k+1\}\^\{G\}$ r_j, si pone i_k ← $\lfloor$rem / prod$\rfloor$ e rem ← rem mod prod. L'impacchettamento è perfettamente reversibile tramite aritmetica modulare, senza alcuna perdita di informazione nel passaggio dalla rappresentazione indici alla rappresentazione token e viceversa.

I vantaggi dell'impacchettamento misto-radix rispetto alle alternative sono molteplici. La reversibilità perfetta garantisce che nessuna informazione venga persa nella tokenizzazione. La compressione è quasi ottimale per i raggi dati, in quanto l'intero spazio dei valori possibili dei token viene utilizzato senza gaps. L'assenza di un codebook apprendibile elimina la dipendenza da strutture di dati aggiuntive e i problemi di codebook collapse. Le operazioni sono esclusivamente su interi, quindi efficienti su hardware dedicato. Il parametro G bilancia la dimensione del vocabolario — 4^G — e il tasso di token — f × $\lceil D/G \rceil$ — permettendo un trade-off controllato in funzione dei requisiti dell'applicazione.

VII.4.3 — Il decoder HiFi-GAN con gate DAAM

Il decoder HiFi-GAN ricostruisce la forma d'onda audio a partire dalle feature quantizzate attraverso una serie di blocchi di upsampling con blocchi residuali a campo recettivo multiplo (MRF). L'architettura decoder trasforma le feature quantizzate $\mathbf{z}_q \in \mathbb{R}\^(B \times 512 \times T_z)$ attraverso blocchi ConvTranspose1D che espandono progressivamente la dimensione temporale e riducono il numero di canali:

512 → 384 → 256 → 128 → 64

con strides 6, 5, 5, 8, 8 (stride totale 9 600), producendo la forma d'onda ricostruita $\hat{\mathbf{x}} \in \mathbb{R}\^(B \times 1 \times T_wav)$. Ciascun blocco decoder comprende una convoluzione trasposta per l'upsampling e blocchi residuali MRF con kernel di

dimensioni diverse [3, 7, 11, 15, 23, 32] per catturare strutture temporali a scale multiple. Le attivazioni SnakeBeta forniscono il bias induttivo periodico appropriato per la sintesi audio ad alta fedeltà. I moduli DAAM vengono applicati (opzionalmente) nei blocchi residuali del decoder per modulare adattivamente le feature durante la ricostruzione.

VII.4.4 — L'obiettivo di addestramento del Secondo Stadio

Il Secondo Stadio ottimizza congiuntamente il quantizzatore FSQ, il decoder HiFi-GAN e l'encoder JEPA (fine-tuned) rispetto alla perdita composita:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{stft} \cdot \mathcal{L}_{stft} + \lambda_{gan} \cdot \mathcal{L}_{gan}$$

Perdita di ricostruzione L1. La perdita di ricostruzione è la media assoluta delle differenze campione per campione tra la forma d'onda ricostruita $\hat{\mathbf{x}}$ e quella originale $\mathbf{x}$:

$$\mathcal{L}_{rec} = (1/T_{wav}) \cdot \Sigma_{t=1}^{T_{wav}} |\hat{x}_t - x_t|$$

La scelta della norma L1 anziché L2 riflette la struttura della distribuzione degli errori di ricostruzione audio, dove valori estremi sono meno rari che in segnali visivi e la L1 è più robusta a questi outlier.

Perdita STFT multi-risoluzione. La perdita spettrale multi-risoluzione combina perdite di convergenza spettrale e di magnitudine logaritmica su M configurazioni STFT diverse (con dimensioni FFT: [2 048, 1 024, 512, 256, 128]; hop sizes: [512, 256, 128, 64, 32]; finestra: Hann):

$$\mathcal{L}_{stft} = \Sigma_{m=1}^{M} (\mathcal{L}^{(m)}_{sc} + \mathcal{L}^{(m)}_{mag})$$

con convergenza spettrale:

$$\mathcal{L}^{(m)}_{sc} = \| |S_m(\hat{\mathbf{x}})| - |S_m(\mathbf{x})| \|_F / \| |S_m(\mathbf{x})| \|_F$$

e perdita di magnitudine logaritmica:

$$\mathcal{L}^{(m)}_{mag} = (1/N_m) \cdot \| \log|S_m(\hat{\mathbf{x}})| - \log|S_m(\mathbf{x})| \|_1$$

dove S_m denota la trasformata STFT con la m-esima configurazione, $\|\cdot\|_F$ è la norma di Frobenius e $\|\cdot\|_1$ è la norma matriciale L1 (somma dei valori assoluti). La perdita multi-risoluzione valuta la fedeltà spettrale a scale temporali multiple, dalla struttura grossolana del segnale (FFT ampia) ai dettagli transitori rapidi (FFT stretta).

Perdita avversariale GAN. Il sistema impiega discriminatori multi-periodo (MPD) e multi-scala (MSD) nello stile di HiFi-GAN. La perdita del generatore è:

$$\mathcal{L}_{gen} = \Sigma_{d \in \{MPD, MSD\}} \mathbb{E}[(D_d(\hat{\mathbf{x}}) - 1)^2]$$

La perdita di feature matching è:

$\mathcal{L}_feat = \Sigma_\{d \in \{MPD, MSD\}\} \Sigma_\{l=1\}^\{L_d\} (1/N_l) \cdot \|D^(l)_d(\mathbf{x}) - D^(l)_d(\hat{\mathbf{x}})\|_1$

La perdita GAN totale è $\mathcal{L}_gan = \mathcal{L}_gen + \mathcal{L}_feat$, e la perdita del discriminatore è:

$\mathcal{L}_disc = \Sigma_\{d \in \{MPD, MSD\}\} (\mathbb{E}[(D_d(\mathbf{x}) - 1)^2] + \mathbb{E}[D_d(\hat{\mathbf{x}})^2])$

I pesi dei termini di perdita sono $\lambda_stft = 2{,}0$ e $\lambda_gan = 0{,}1$, con un periodo di riscaldamento di 5 000 passi durante i quali il discriminatore è congelato e soltanto la perdita di ricostruzione e quella spettrale sono attive. Dopo il riscaldamento il discriminatore viene aggiornato ad ogni passo.

Gli iperparametri del Secondo Stadio sono: ottimizzatore AdamW con $\beta_1 = 0{,}8$, $\beta_2 = 0{,}99$; tasso di apprendimento $1{,}5 \times 10^\{-4\}$ per il decoder e $0{,}75 \times 10^\{-4\}$ per i discriminatori; weight decay $10^\{-3\}$; batch size 8; 29 000 passi di addestramento.

VII.5 — Configurazione sperimentale e architettura complessiva

L'intero sistema viene addestrato su LibriLight, un corpus di parlato in inglese non etichettato di circa 9 000 ore complessivamente distribuito fra i due stadi. Il segnale audio viene ricampionato a 24 kHz in forma monoaurale senza ulteriore preprocessing, con la normalizzazione gestita internamente dall'architettura del modello. L'addestramento distribuito utilizza due GPU NVIDIA A100 da 80 GB con precisione mista FP16 per le operazioni di forward e backward e FP32 per le operazioni critiche.

Il parametro complessivo del sistema è il seguente: l'encoder online del Primo Stadio conta 121,7 milioni di parametri addestrabili; l'encoder target (EMA) ne ha 118,5 milioni aggiornati per momentum; la rete predittiva 3,2 milioni addestrabili. Il Secondo Stadio aggiunge il quantizzatore FSQ (circa 0,01 milioni di parametri addestrabili) e il decoder HiFi-GAN (69,2 milioni addestrabili). Il modello finale per l'inferenza — comprendente solo l'encoder online, il quantizzatore FSQ e il decoder HiFi-GAN, senza la rete predittiva né il target encoder — conta 191,0 milioni di parametri in un singolo forward pass.

VII.6 — Discussione: perché DAAM migliora le rappresentazioni JEPA per il parlato

La riduzione del 47% nella perdita MSE del Primo Stadio ottenuta da JEPA+DAAM rispetto a JEPA standard è un risultato notevole che merita un'analisi teorica più approfondita. La domanda naturale è: attraverso quale meccanismo il gate DAAM produce questo miglioramento? La risposta non è ovvia perché DAAM non modifica la struttura del compito predittivo — le stesse regioni vengono mascherate, lo stesso predictor deve predirle — ma soltanto la weighting delle feature dell'encoder.

La spiegazione più convincente è geometrica. Nel framework della predizione latente, il compito del predictor è trovare una trasformazione che mappi le rappresentazioni del contesto alle rappresentazioni del target. Se le rappresentazioni del contesto e del target sono nello stesso spazio semantico — se misurano le stesse proprietà del segnale, differenziandosi soltanto per quale regione temporale descrivono — allora la trasformazione richiesta al predictor è relativamente semplice: una sorta di "interpolazione strutturata" nello spazio latente. Se invece le rappresentazioni del contesto sono distorte rispetto a quelle del target — se enfatizzano caratteristiche diverse da quelle che il target encoder enfatizza — allora il predictor deve apprendere una correzione della distorsione prima di poter fare la predizione effettiva.

DAAM riduce questa distorsione selezionando, attraverso il gate gaussiano, le feature temporalmente più salienti — quelle che il target encoder enfatizzerà — e amplificandole nel context encoder prima della predizione. Il risultato è che le rappresentazioni del contesto sono "pre-allineate" con quelle del target nello spazio semantico, rendendo il compito del predictor più diretto. Questa interpretazione è coerente con l'osservazione che DAAM produce convergenza più rapida (non solo perdita finale inferiore): l'allineamento pre-computato riduce il numero di step necessari per trovare la trasformazione predittiva corretta.

I risultati empirici mostrano che JEPA+DAAM converge a una perdita MSE del Primo Stadio di circa 0,09

I risultati empirici mostrano che JEPA+DAAM converge a una perdita MSE del Primo Stadio di circa 0,09, contro circa 0,17 di JEPA senza DAAM, con convergenza più rapida e stabile. Questa differenza di quasi il cinquanta per cento nella perdita finale ha un'interpretazione teorica che merita di essere sviluppata.

L'attenzione softmax standard risponde alla domanda delle correlazioni pairwise: dato il fotogramma t, quali altri fotogrammi temporali gli sono più simili? Questo tipo di attenzione è efficace per catturare dipendenze di contenuto ma è computazionalmente costoso ($O(T^2)$) e non distingue intrinsecamente le regioni statisticamente salienti da quelle statisticamente ordinarie della sequenza. DAAM invece calcola la salienza statistica di ciascun fotogramma in modo indipendente: la densità di un fotogramma sotto la mistura gaussiana appresa misura quanto quel fotogramma è "tipico" rispetto alla distribuzione temporale delle feature — con gate elevato per fotogrammi statisticamente rappresentativi e gate basso per fotogrammi anomali o transitori. Nel dominio del parlato, le transizioni fonemiche, gli attacchi consonantici e i nuclei vocalici corrispondono a regioni dove il segnale mostra alta densità sotto la distribuzione gaussiana delle sue statistiche locali — sono le regioni più "tipiche" e informativamente più dense — e il gate DAAM amplifica selettivamente le loro feature, rendendo l'encoder più sensibile alle strutture foneticamente rilevanti.

Questa modalità di selezione adattiva delle feature è particolarmente appropriata per il compito JEPA di predizione latente con mascheramento a blocchi: le regioni

mascherate che devono essere predette contengono tipicamente strutture fonetiche coerenti, e un encoder che identifica e amplifica le feature statisticamente salienti nel contesto è meglio equipaggiato per produrre rappresentazioni del contesto da cui il predictor può estrarre informazioni predittive rilevanti.

La struttura gerarchica dello spazio delle rappresentazioni — conseguenza del downsampling progressivo da strato a strato dell'encoder convoluzionale — produce, in combinazione con DAAM, una scoperta automatica della struttura fonica: i livelli inferiori dell'encoder catturano feature spettrali locali, i livelli superiori integrano queste feature in strutture fonemiche e sillabiche, e il meccanismo DAAM ai vari livelli seleziona adattivamente le regioni temporali più informative a ciascuna scala. La compressione a 2,5 Hz — un fattore 9 600 rispetto alla frequenza di campionamento originale — non è arbitraria ma corrisponde approssimativamente alla frequenza di sillabe nel parlato lento (2–3 sillabe al secondo), suggerendo che l'encoder JEPA con DAAM scopre organicamente la struttura sillabica del segnale vocale come struttura gerarchica fondamentale.

Capitolo VIII — L'Estensione Multimodale del Framework AMI: JEPA come Tokenizzatore Neurale Universale e la Struttura della Percezione Incarnata

VIII.1 — Il parlato come banco di prova per i principi generalizzati di JEPA

Il Capitolo VII ha sviluppato nel dettaglio l'applicazione del principio JEPA al dominio del parlato attraverso l'architettura a due stadi con attenzione adattiva di densità e quantizzazione scalare finita. Prima di procedere alla discussione delle connessioni di questa architettura con il framework AMI e con i principi teorici sviluppati nei capitoli precedenti, è utile porre esplicitamente la domanda metodologica che sottende l'intero argomento: il fatto che il principio JEPA — predizione latente con encoder target EMA, senza ricostruzione sensoriale nel primo stadio — funzioni per le immagini, per i video e ora anche per il parlato è una coincidenza empirica o riflette una proprietà generale del principio?

La risposta suggerita dall'analisi teorica dei capitoli precedenti è che si tratta di una proprietà di principio. Il meccanismo fondamentale con cui JEPA evita le patologie dei paradigmi generativi a livello sensoriale — la multimodalità della distribuzione del target nello spazio delle osservazioni e il bias verso caratteristiche locali a bassa frequenza a scapito delle strutture semantiche globali — è intrinseco alla struttura dell'obiettivo di apprendimento, non alle caratteristiche specifiche del dominio modello. In qualsiasi dominio in cui le osservazioni sensoriali siano generate da un processo che opera a molteplici livelli di astrazione — caratteristiche fisiche di basso livello e strutture semantiche di alto livello — la predizione nello spazio latente produrrà rappresentazioni qualitativamente migliori di quelle apprese

attraverso la ricostruzione delle osservazioni grezze, per la semplice ragione che la distribuzione condizionale target nello spazio latente è intrinsecamente meno multimodale di quella nello spazio delle osservazioni.

La conferma sperimentale di questo argomento per tre domini distinti — immagini statiche, sequenze video, forme d'onda audio — suggerisce con forza che JEPA è un principio di apprendimento delle rappresentazioni ad ampio spettro modale, non un'architettura specializzata per la visione. Questa ipotesi di generalità ha implicazioni profonde per il framework AMI, dove il Perception Module deve integrare segnali da molteplici canali sensoriali in una rappresentazione unificata del mondo: se JEPA è applicabile a ciascun canale modalità, allora il Perception Module può essere costruito come una famiglia di encoder JEPA modali separati le cui rappresentazioni vengono poi integrate a livello superiore, preservando i vantaggi teorici del principio in ciascun dominio.

VIII.2 — Il tasso di fotogrammi come scelta teorica, non ingegneristica

Una delle decisioni architetturali più significative del sistema descritto nel Capitolo VII è la scelta del tasso di fotogrammi di 2,5 Hz, ottenuto attraverso un downsampling complessivo di 9 600 campioni per hop a 24 000 Hz. Questo valore è sufficientemente basso da ridurre il costo computazionale dell'elaborazione e della tokenizzazione — 47,5 token al secondo contro i 75 Hz di SoundStream e EnCodec — ma sufficientemente alto da non perdere strutture fonetiche essenziali.

Il confronto con i codec neurali esistenti rivela la gerarchia dei tassi di fotogramma nel panorama attuale della compressione audio neurale: SoundStream e EnCodec operano a 75 Hz (13,3 ms per fotogramma), DAC a 86 Hz, DualCodec nell'intervallo 12,5–25 Hz, Mimi a 12,5 Hz, U-Codec a 5 Hz, mentre il sistema JEPA+DAAM descritto nel Capitolo VII raggiunge 2,5 Hz — il tasso più basso fra tutti i sistemi comparabili. Questo posizionamento all'estremo inferiore dello spettro dei tassi di fotogramma non è accidentale: il principio JEPA, applicando il downsampling aggressivo attraverso l'architettura convoluzionale strided, costruisce il sistema in modo che la compressione avvenga principalmente nell'encoder — prima dell'obiettivo di predizione latente — piuttosto che nel quantizzatore. Le rappresentazioni a 2,5 Hz non sono il risultato di una quantizzazione lossy di feature a frequenza più alta, ma sono le rappresentazioni native apprese dall'encoder JEPA con mascheramento a blocchi, il che significa che l'informazione non presente a questa scala temporale non era semanticamente necessaria per il compito predittivo.

Questa osservazione suggerisce un principio di progettazione generale per i codec neurali basati su apprendimento autosupervisionato: il tasso di fotogramma ottimale non è determinato dalla fedeltà della ricostruzione sensoriale — che scala sempre in favore di tassi più alti — ma dalla granularità temporale delle strutture semantiche che si desidera catturare. Per il parlato, questa granularità è

approssimativamente quella sillabica (2–3 fotogrammi al secondo), coerente con il tasso di 2,5 Hz adottato. Per applicazioni che richiedono la cattura di strutture fonemiche sottili a scala infra-sillabica, un tasso leggermente più alto potrebbe essere preferibile; per applicazioni che richiedono solo la comprensione semantica a livello di parola o di frase, il tasso di 2,5 Hz potrebbe essere addirittura superiore al necessario.

Il collegamento con il framework AMI è diretto: il Perception Module di AMI non ha l'obiettivo di costruire una fedele rappresentazione sensoriale del mondo ma di produrre un'astrazione sufficiente a supportare la pianificazione e il ragionamento. Per il canale uditivo, questa astrazione dovrebbe catturare il contenuto linguistico e prosodico del segnale vocale ma non i dettagli acustici irrilevanti per il significato. Il tasso di 2,5 Hz sembra essere un buon compromesso in questo senso, e il fatto che emerga organicamente dall'ottimizzazione JEPA — senza essere prescritto a priori — costituisce un'evidenza a favore della tesi che il principio JEPA apprende intrinsecamente le strutture temporali semanticamente rilevanti al giusto livello di astrazione.

VIII.3 — La mistura gaussiana come modello della distribuzione temporale del segnale vocale

Il meccanismo DAAM si distingue dalle forme di attenzione adottate nei sistemi JEPA per immagini e video per un aspetto concettuale fondamentale: mentre l'attenzione nei Vision Transformer opera nello spazio delle correlazioni tra patch (una relazione di secondo ordine tra le rappresentazioni), DAAM opera nello spazio della probabilità delle feature (una relazione di ordine zero, ovvero una proprietà intrinseca di ciascuna posizione temporale indipendentemente dalle altre). Questa differenza ha conseguenze sulla natura delle strutture che il meccanismo può apprendere a identificare come salienti.

La mistura gaussiana con $K = 4$ componenti appresa da DAAM non è una rappresentazione esplicita della fonetica o della prosodia, ma è un modello implicito della distribuzione marginale delle feature estratte dall'encoder lungo la dimensione temporale. Durante l'addestramento, i parametri della mistura — gli offset di media δ_k e i parametri di log-scala ν_k — vengono ottimizzati congiuntamente con tutti gli altri pesi dell'encoder rispetto all'obiettivo di predizione JEPA. I valori a convergenza rifletteranno quindi la struttura statistica delle feature che l'encoder ha imparato a estrarre, che a sua volta riflette la struttura statistica del parlato necessaria per la predizione latente.

L'interpretazione geometrica del gate DAAM è illuminante. A partire dalle statistiche temporali μ e σ^2 computate su ciascuna feature map, il gate $G(\mathbf{x}_t)$ assegna un valore più alto ai fotogrammi la cui deviazione dalla media temporale è più coerente con le gaussiane apprese. Fotogrammi con valori estremi in alcune feature — corrispondenti a transizioni fonetiche brusche, a picchi di energia consonantica o a variazioni rapide di formante — avranno deviazioni dalla media temporale elevate e potranno ricevere gate alti o bassi a seconda di quale gaussiana

della mistura cattura quelle deviazioni come tipiche. La molteplicità dei componenti K permette alla mistura di modellare una distribuzione bimodale o multimodale delle deviazioni temporali, catturando il fatto che il segnale vocale ha tipicamente non uno ma diversi "stati" statisticamente distinti nell'arco di un enunciato.

L'affinità concettuale con i modelli Hidden Markov del parlato — dove il segnale vocale viene modellato come un processo che transita tra stati latenti gaussiani — è evidente, anche se il DAAM non ha la struttura Markoviana esplicita degli HMM. In DAAM, la mistura gaussiana modella la distribuzione marginale delle feature in modo soft e differenziabile, senza fare assunzioni sull'ordine temporale degli stati. Questo è sia un punto di forza — evita le restrizioni parametriche degli HMM — sia un punto di debolezza rispetto alla modellazione esplicita delle transizioni: il gate di ciascun fotogramma è calcolato indipendentemente dagli altri, senza tener conto della sequenza degli stati precedenti e successivi.

VIII.4 — JEPA come tokenizzatore neurale universale: la connessione con i modelli linguistici di grandi dimensioni

Una delle motivazioni esplicite del sistema descritto nel Capitolo VII è la produzione di token discreti adatti all'ingestione da parte di modelli linguistici di grandi dimensioni. Il tasso di 47,5 token al secondo a un vocabolario di 16 384 elementi pone questo sistema in una posizione peculiare nell'ecosistema dei tokenizzatori audio per LLM.

Il vocabolario di 16 384 elementi è paragonabile al vocabolario di tokenizzatori subword adoperati in alcuni LLM di dimensione media, il che significa che i token audio prodotti dal sistema JEPA+FSQ possono in linea di principio essere intercalati con token testuali in un unico transformer autoregressivo senza richiedere vocabolari separati o meccanismi speciali di codifica della modalità. Questo è esattamente lo schema adottato in sistemi di speech language model come SpeechGPT, AudioPaLM e altri, dove la generazione audio viene trattata come generazione di token discreti in un vocabolario allargato che comprende sia token testuali che token audio.

Il tasso di 47,5 token al secondo è sufficientemente basso da rendere computazionalmente gestibile la generazione autoregressiva di sequenze audio di durata ordinaria: un enunciato di 5 secondi corrisponde a circa 237 token, una sequenza breve per qualsiasi LLM moderno. A titolo di confronto, a 75 Hz (EnCodec con 8 codebook) la stessa sequenza corrisponderebbe a $375 \times 8 = 3\,000$ token — quasi 13 volte di più. Questa differenza quantitativa ha implicazioni qualitative per la fattibilità dell'integrazione audio-testo in un unico modello autoregressivo: a 47,5 token al secondo la proporzione di contesto occupato dall'audio rispetto al testo rimane gestibile anche per conversazioni di durata considerevole, mentre a 75 Hz o superiore l'audio rischia di dominare il contesto del transformer lasciando spazio insufficiente all'elaborazione del testo.

La reversibilità dell'impacchettamento misto-radix garantisce che dal token sia sempre possibile ricavare esattamente gli indici FSQ originali, e da questi le feature quantizzate, e da queste la forma d'onda ricostruita tramite il decoder HiFi-GAN. La catena di decodifica è dunque: token → indici FSQ (decodifica misto-radix) → feature quantizzate (dequantizzazione FSQ) → forma d'onda ricostruita (HiFi-GAN decoder). Ogni passo è deterministico e privo di perdita di informazione fino all'ultimo — la ricostruzione HiFi-GAN introduce inevitabilmente una perdita rispetto all'audio originale, ma tutta l'informazione semanticamente rilevante catturata dall'encoder JEPA è preservata nei token.

Nell'architettura AMI descritta nel Capitolo V, questo schema di tokenizzazione apre la possibilità di un Perception Module che produce non solo rappresentazioni continue delle osservazioni sensoriali — come descritto nella sezione V.2 — ma anche sequenze di token discreti direttamente utilizzabili come input per il World Model implementato come un transformer autoregressivo. La distinzione fra la rappresentazione continua (per la pianificazione e il ragionamento nello spazio latente) e la rappresentazione discreta (per la generazione e la comprensione del linguaggio) non è necessariamente una scelta architetturale rigida: il framework JEPA+FSQ produce entrambe — le feature continue $\mathbf{z}_q$ dopo la dequantizzazione e i token interi compatti — dalla medesima pipeline di elaborazione.

VIII.5 — Limitazioni e direzioni di sviluppo futuro

Un'analisi obiettiva del sistema descritto nel Capitolo VII deve riconoscere le sue limitazioni strutturali, che emergono chiaramente dalla lettura del documento originale.

La strategia di mascheramento a blocchi con distribuzioni di span fisse potrebbe non adattarsi in modo ottimale alle variazioni del ritmo e della struttura linguistica presenti in parlanti diversi, stili del discorso diversi o lingue diverse. Il sistema è stato sviluppato e valutato esclusivamente su parlato in inglese (LibriLight), e la sua applicabilità a lingue tonali — dove l'altezza vocalica porta contenuto lessicale — o a lingue morfologicamente ricche — dove le strutture di sub-parola sono più complesse — rimane una questione aperta. Un parlato in lingua mandarino, per esempio, richiederebbe probabilmente un tasso di fotogrammi superiore a 2,5 Hz per catturare adeguatamente la struttura tonale, e la strategia di mascheramento dovrebbe essere sensibile alle unità tonali piuttosto che ai blocchi temporali uniformi.

Il pretraining è stato condotto su una quantità di dati relativamente modesta rispetto ai sistemi di auto-supervisione allo stato dell'arte — circa 9 000 ore contro le centinaia di migliaia di ore adoperate da WavLM e HuBERT — e la valutazione è qualitativa piuttosto che quantitativa su benchmark standardizzati. Questo rende difficile la collocazione precisa del sistema nell'ecosistema delle rappresentazioni vocali autosupervisionato e lascia aperta la domanda su come le prestazioni scalino con il volume dei dati di pretraining.

La valutazione della qualità di ricostruzione del decoder HiFi-GAN non è riportata attraverso metriche standard come PESQ, STOI o MOS, il che rende impossibile confrontare obiettivamente la fedeltà acustica del sistema con quella dei codec neurali esistenti. Sebbene il tasso di token molto basso (47,5 token/s contro i 75 Hz di EnCodec) suggerisca inevitabilmente una riduzione della fedeltà rispetto alla ricostruzione ad alta risoluzione, la natura semanticamente densa delle rappresentazioni JEPA potrebbe compensare parzialmente questa perdita per compiti di comprensione e generazione del parlato — ma questa ipotesi richiede verifica empirica quantitativa.

L'utilizzo di un numero fisso K = 4 di componenti gaussiane nel DAAM è una scelta pragmatica non accompagnata da una giustificazione teorica della scelta del valore specifico. È plausibile che il numero ottimale di componenti vari con la complessità della distribuzione delle feature — e quindi con l'architettura dell'encoder e con il tipo di parlato — e che un meccanismo adattivo che apprende K o seleziona automaticamente il numero di componenti attivi produca rappresentazioni di qualità superiore.

VIII.6 — La struttura duale continua-discreta come contributo architetturale al framework AMI

Uno degli aspetti meno discussi nel documento originale ma potenzialmente più significativi per il framework AMI è la produzione sistematica di rappresentazioni in due forme strutturalmente diverse ma derivate dalla medesima pipeline: le feature continue **z** a 2,5 Hz prodotte dall'encoder JEPA (utilizzabili direttamente come input per il World Model o per compiti di riconoscimento e classificazione) e i token discreti a 47,5 token al secondo (utilizzabili come input per modelli linguistici autoregressivi o per la generazione controllata di parlato).

Questa dualità riflette una tensione fondamentale nell'architettura del Perception Module di AMI: il World Model ha bisogno di rappresentazioni continue per la simulazione e la pianificazione nello spazio latente, ma i modelli linguistici — che nel framework AMI svolgono un ruolo nel ragionamento di alto livello — operano su vocabolari discreti. La soluzione tradizionale è di mantenere due encoder separati, uno per ciascuna rappresentazione, con tutti i costi computazionali e le difficoltà di allineamento che ne conseguono. Il framework JEPA+FSQ suggerisce invece una soluzione unificata: un singolo encoder apprende la rappresentazione continua ottimale per il compito predittivo, e il quantizzatore FSQ ne fornisce una versione discreta reversibile senza costi aggiuntivi di addestramento significativi.

La connessione con il principio teorico di LeJEPA sviluppato nella sezione III.5bis è qui particolarmente evidente. Se le rappresentazioni continue prodotte dall'encoder JEPA per il parlato seguono la distribuzione gaussiana isotrópica ottimale dimostrata dal Teorema 1 di LeJEPA — ipotesi che sarebbe utile verificare empiricamente per questo dominio — allora la quantizzazione FSQ di queste rappresentazioni produce token che sono, in senso informativo, proiezioni ottimali della distribuzione gaussiana isotrópica su una griglia discreta finita. La

struttura dell'impacchettamento misto-radix garantisce poi che questi token siano decodificabili in modo esatto, preservando l'intera informazione della rappresentazione discreta. Il percorso token → feature quantizzate → feature continue (tramite interpolazione o dequantizzazione morbida) dovrebbe quindi produrre rappresentazioni continue che si avvicinano alla distribuzione gaussiana isotrópica originale, con una distorsione controllata dalla granularità della quantizzazione.

VIII.7 — Epilogo: verso una teoria unificata dell'apprendimento delle rappresentazioni multimodali

I capitoli precedenti di questa monografia hanno sviluppato, attraverso il percorso che va dalle fondamenta epistemologiche del Capitolo I alle implicazioni computazionali del Capitolo VI, una tesi centrale: che il paradigma dominante dell'apprendimento autosupervisionato — basato sulla predizione o ricostruzione nello spazio delle osservazioni sensoriali — sia strutturalmente inadeguato alla costruzione di rappresentazioni del mondo sufficientemente astratte, stabili e composizionali da supportare la pianificazione e il ragionamento di un agente autonomo. JEPA e le sue estensioni — I-JEPA, V-JEPA, MC-JEPA, LeJEPA, e ora il sistema JEPA+DAAM per il parlato — rappresentano una risposta coerente a questa inadeguatezza strutturale, basata sul principio della predizione nello spazio latente come obiettivo intrinseco.

I contributi teorici più recenti analizzati nelle sezioni integrative — LeJEPA per la caratterizzazione della distribuzione ottimale delle rappresentazioni, la Semantic Tube Prediction per la regolarizzazione geodetica delle traiettorie latenti — forniscono per la prima volta una base normativa che trasforma JEPA da architettura empiricamente vincente a sistema teoricamente derivato da principi di ottimalità. Il Teorema 1 di LeJEPA afferma che la distribuzione gaussiana isotrópica è il minimizzatore del rischio downstream per qualsiasi compito di probing su embedding vincolati scalari sulla covarianza; la Proposizione 2.1 di STP afferma che le traiettorie dei token a convergenza seguono equazioni differenziali ordinarie il cui perturbazione stocastica produce coni di inferenza di raggio proporzionale alla radice quadrata del tempo. Questi risultati, pur derivati per domini specifici — rispettivamente immagini e testo — hanno la struttura matematica di teoremi di tipo "ottimalità della gaussiana" e "principio di minima azione" che trascendono le specificità del dominio.

L'estensione di questi principi al parlato attraverso il sistema JEPA+DAAM aggiunge un'ulteriore dimensione alla teoria: il meccanismo DAAM introduce, per la prima volta in un'architettura JEPA, una forma di attenzione basata sulla probabilità distribuzionale piuttosto che sulla correlazione pairwise. La congettura teorica — che non viene formalizzata nel documento originale ma emerge naturalmente dall'analisi — è che DAAM approssima, in modo differenziabile e computazionalmente efficiente, una forma di selezione adattiva delle feature che massimizza l'informazione statistica delle rappresentazioni del contesto rispetto al

target mascherato. Se questa congettura fosse formalizzata e dimostrata, il DAAM potrebbe essere visto come un'istanza di un principio più generale: che l'attenzione in architetture JEPA dovrebbe operare non sulle correlazioni tra feature ma sulla loro saliezza distribuzionale — un principio che potrebbe guidare lo sviluppo di meccanismi di attenzione per altri domini modali dove la struttura statistica è più ricca della struttura correlazionale.

La teoria unificata dell'apprendimento delle rappresentazioni multimodali verso cui questi sviluppi convergono avrebbe la seguente struttura. Per ciascun dominio modale d con spazio delle osservazioni $\mathcal{X}_d$, esiste un encoder ottimale $\varphi_d : \mathcal{X}_d \to \mathbb{R}^{K_d}$ tale che la distribuzione marginale delle rappresentazioni è gaussiana isotrópica (per il Teorema 1 di LeJEPA), le rappresentazioni soddisfano un criterio di predizione latente basato sul mascheramento (per il principio JEPA), e le traiettorie delle rappresentazioni nel tempo seguono geodetiche nel manifold latente (per la Geodesica Hypothesis di STP). Il Perception Module di AMI è la composizione di questi encoder modalità-specifici con un modulo di integrazione che produce una rappresentazione unificata dello stato del mondo nello spazio comune del World Model.

La distanza da questa teoria unificata alla sua realizzazione pratica è ancora considerevole. Per il dominio visivo, I-JEPA e LeJEPA forniscono la componente predittiva e la componente distribuzionale; STP fornisce la componente geodetica per le sequenze; ma l'integrazione dei tre principi in un sistema unificato e la loro applicazione coordinata a tutti i domini modali rilevanti per AMI — visione, udito, tatto, propriocezione, linguaggio — rimane un programma di ricerca aperto. Il sistema JEPA+DAAM per il parlato descritto nel Capitolo VII è, in questo senso, un passo significativo nella direzione giusta: dimostra che il principio JEPA si estende naturalmente a un dominio modale completamente diverso da quello per cui era stato originariamente concepito, e che l'integrazione di meccanismi di attenzione basati sulla probabilità distribuzionale — come DAAM — può migliorare la qualità delle rappresentazioni apprese riducendo la perdita del Primo Stadio di quasi il cinquanta per cento.

L'argomentazione epistemologica del Capitolo I aveva identificato come radice della crisi del paradigma generativo a livello sensoriale il fatto che gli obiettivi di apprendimento dominanti non distinguono fra strutture semantiche e varianti percettive irrilevanti. I sistemi analizzati in questa monografia — da I-JEPA a LeJEPA a STP a JEPA+DAAM — convergono tutti verso la medesima soluzione: spostare il bersaglio dell'apprendimento dallo spazio delle osservazioni grezze allo spazio delle strutture semantiche astratte, garantendo che le rappresentazioni apprese siano semanticamente informative attraverso vincoli geometrici e distribuzionali teoricamente motivati. La convergenza di risultati empirici positivi su domini modali multipli — immagini, video, parlato — con giustificazioni teoriche sempre più solide suggerisce che questa soluzione non sia contingente ma risponda a una proprietà fondamentale della struttura dell'informazione nei segnali naturali: che i contenuti semanticamente rilevanti siano intrinsecamente più regolari, più compressibili e più predittibili delle varianti sensoriali irrilevanti, e che

un encoder progettato per predire il primo debba per necessità imparare a ignorare le seconde.

Capitolo IX — Comprensione, Coscienza e Causalità: Una Critica Filosofica delle Architetture Predittive come Candidati all'Intelligenza Generale

IX.1 — Il problema che la tecnica non può risolvere da sola

Prima di avventurarsi nell'analisi dell'argomento della Stanza Cinese, è utile precisare perché queste questioni filosofiche abbiano rilevanza pratica per l'ingegneria dei sistemi intelligenti — una rilevanza che va oltre il dominio delle discussioni accademiche e tocca direttamente le decisioni di design architetturale. Se la comprensione richiede radicamento causale nel mondo fisico (come argomentano Harnad e la tradizione dell'embodied cognition), allora l'addestramento su dati passivi — immagini, testo, video — non può in linea di principio produrre comprensione nel senso rilevante, indipendentemente dalla quantità di dati o dalla sofisticazione dell'architettura. Se la semantica non è riducibile alla sintassi (come sostengono Searle, Putnam e Tarski attraverso argomenti diversi), allora nessun sistema formale — per quanto complesso — può acquisire contenuto intenzionale attraverso operazioni formali. Se la sistematicità cognitiva richiede strutture composizionali esplicite (come argomentano Fodor e Pylyshyn), allora rappresentazioni distribuite — per quanto semanticamente ricche — non sono sufficienti per la generalizzazione composizionale.

Queste non sono speculazioni filosofiche di maniera: sono vincoli strutturali che, se validi, delimitano in modo preciso ciò che sistemi come JEPA e AMI possono e non possono conseguire. Ignorarli non li fa scomparire; li rende semplicemente invisibili — il che è peggio, perché impedisce di progettare sistemi consapevoli dei propri limiti.

Nei capitoli precedenti la monografia ha sviluppato con rigore crescente la tesi che il paradigma JEPA, nelle sue varianti teoricamente fondate — I-JEPA, LeJEPA, JEPA+DAAM — costituisce un progresso strutturale rispetto ai paradigmi generativi a livello sensoriale nell'apprendimento di rappresentazioni semanticamente ricche, e che il framework AMI di LeCun offre una visione architettonica coerente per l'integrazione di tali rappresentazioni in un sistema capace di pianificazione e ragionamento. Le evidenze empiriche — la convergenza a perdite di predizione inferiori, la violazione delle Scaling Laws attraverso STP, le prestazioni di LeJEPA su dataset specializzati — forniscono sostegno sperimentale a questa tesi. Ma vi è un piano sul quale nessun benchmark, per quanto ampio, e nessun teorema, per quanto elegante, può rispondere alle domande più urgenti sull'intelligenza artificiale generale: il piano della comprensione, dell'intenzionalità e della coscienza.

Queste non sono questioni filosofiche nel senso di questioni oziose o decorative rispetto all'ingegneria reale. Sono questioni che toccano il nucleo di ciò che si intende per AGI nel senso in cui questa espressione è usata dagli autori stessi citati in questa monografia. LeCun è esplicito nel sostenere che i sistemi attuali — inclusi i più capaci LLM — mancano di comprensione nel senso rilevante, e che il suo framework AMI aspira a sistemi che comprendano il mondo in modo non superficiale. Ma cosa significa "comprendere" in un senso che non sia ridotto a "generalizzare bene su benchmark"? E può un sistema che apprende attraverso la minimizzazione di perdite di predizione — per quanto geometricamente ben strutturate — aspirare a quella forma di comprensione? Queste domande non hanno risposte consensuali nella filosofia della mente contemporanea, e onestà intellettuale richiede di non fingere che esistano.

IX.2 — L'argomento della Stanza Cinese e le sue conseguenze per i sistemi predittivi

John Searle pubblicò nel 1980 sull'importante rivista Behavioral and Brain Sciences un esperimento mentale destinato a ridefinire il dibattito sull'intelligenza artificiale per i decenni successivi. L'argomento è noto come la Stanza Cinese (Chinese Room). Immaginiamo un parlante esclusivo di inglese chiuso in una stanza, armato di un manuale di regole sintattiche per la manipolazione di simboli cinesi. Dall'esterno vengono inseriti foglietti con simboli cinesi (domande); la persona all'interno, seguendo il manuale, produce risposte in simboli cinesi che a un osservatore esterno risultano appropriate e pertinenti. Il sistema — persona più manuale — supera il Test di Turing per la comprensione del cinese. Tuttavia, la persona all'interno non capisce una sola parola di cinese: manipola simboli formali secondo regole sintattiche senza alcuna comprensione del loro significato.

L'argomento di Searle è diretto contro il funzionalismo computazionale nell'accezione forte — la tesi che la mente sia un programma e che qualsiasi sistema fisico che implementa il programma corretto abbia stati mentali genuini, inclusa la comprensione. La Stanza Cinese mostra, secondo Searle, che la sintassi non è sufficiente per la semantica: la corretta manipolazione formale dei simboli non garantisce che i simboli abbiano significato per il sistema che li manipola. La mente biologica non è, secondo Searle, un programma in esecuzione su hardware neurale ma un fenomeno causalmente prodotto dall'architettura specifica del cervello biologico, con proprietà intenzionali — il fatto di essere "di" o "su" qualcosa nel mondo — che non sono implementabili in qualsiasi substrato fisico.

La risposta più comune all'argomento di Searle è la risposta del sistema (systems reply): non è la persona nella stanza a capire il cinese, ma il sistema complessivo — persona più manuale più stanza — che lo capisce in virtù della sua organizzazione funzionale globale. Searle respinge questa risposta notando che se la persona memorizzasse l'intero manuale, l'intera stanza sarebbe "dentro la testa" della persona, e tuttavia la persona continuerebbe a non capire il cinese. Questo controargomento ha a sua volta ricevuto risposte (la risposta del robot, la risposta

del cervello, la risposta di altre menti), e il dibattito è proseguito per decenni senza una conclusione definitiva.

Cosa implica l'argomento di Searle per i sistemi JEPA e per il framework AMI? La risposta dipende da come si interpreta il principio JEPA rispetto alla distinzione sintassi-semantica. Un sistema JEPA che minimizza la perdita di predizione latente manipola strutture formali — vettori in spazi latenti — secondo procedure computazionali determinate dai pesi della rete neurale, senza alcuna garanzia che queste strutture abbiano contenuto intenzionale nel senso di Searle. Il fatto che le rappresentazioni siano "semanticamente ricche" nel senso operativo — ovvero che separino categorie semanticamente distinte nello spazio latente, che catturino relazioni strutturali, che supportino buone prestazioni su benchmark di ragionamento — non implica che il sistema abbia stati mentali nel senso di stati con contenuto intenzionale genuino.

Tuttavia, la critica di Searle si applica con eguale forza a qualsiasi sistema computazionale, incluso il cervello biologico se lo si descrive al livello di elaborazione del segnale neurale. Se la manipolazione formale di simboli è insufficiente per la semantica, e se il cervello biologico è anch'esso, a un certo livello di descrizione, un sistema che manipola formalmente strutture fisiche (potenziali d'azione, pattern di attivazione sinaptica), allora la posizione di Searle — che la semantica richieda la biologia — è difficilmente difendibile senza cadere in una forma di vitalismo computazionale. Il punto non è risolvere il dibattito ma riconoscere che nessuna architettura neurale artificiale — per quanto sofisticata — ha ancora risposto all'argomento in modo soddisfacente.

IX.3 — Il Problema Difficile della Coscienza e il divario esplicativo

Il Problema Difficile della Coscienza occupa una posizione peculiare nel panorama delle questioni filosofiche: è insieme una delle domande più antiche dell'umanità e una delle più recenti acquisizioni della filosofia analitica, nel senso che la sua formulazione precisa e rigorosa è relativamente recente. Le domande sull'anima, sull'interiorità, sul soggetto dell'esperienza attraversano tutta la storia della filosofia — da Platone alla scolastica, da Descartes a Kant, da Hegel a Husserl — ma la formulazione di Chalmers ha il merito di separare con chirurgica precisione le questioni empiricamente indagabili (i Problemi Facili) dalle questioni filosoficamente irriducibili (il Problema Difficile).

La distinzione di Chalmers ha generato un dibattito straordinariamente prolifico. Da un lato, quella che si potrebbe chiamare la "coalizione deflazionista" — che include Dennett, Frankish, Tye e altri — sostiene che il Problema Difficile è un'illusione filosofica, un residuo di intuizioni cartesiane sulla separazione mente-corpo che scompaiono con un'analisi più attenta. Il Problema Difficile sembra difficile, secondo questi autori, non perché ci sia un fatto sul mondo che le teorie funzionaliste non catturano, ma perché i nostri concetti ordinari di

esperienza e qualia sono costruiti in modo tale da sembrare irriducibili a qualsiasi spiegazione fisica — una proprietà dei concetti, non della realtà che descrivono.

Dall'altro lato, la "coalizione realista" — che include oltre a Chalmers anche Nagel, Jackson, Levine e in forme diverse Searle — sostiene che il Problema Difficile è genuino e che nessuna teoria funzionalista può in linea di principio rispondere alla domanda "perché c'è qualcosa che è come essere questo sistema?". L'argomento più potente di questa coalizione è il cosiddetto "zombie argument": sembra concettualmente possibile un essere fisicamente identico a me — con gli stessi pattern neurali, le stesse risposte comportamentali, gli stessi processi computazionali — ma privo di qualsiasi esperienza soggettiva. Se questi "zombie filosofici" sono concettualmente possibili, allora l'esperienza soggettiva non è identica ai processi fisici che la realizzano, e qualsiasi teoria che identifichi le due cose è necessariamente incompleta.

David Chalmers ha articolato, in una serie di lavori dagli anni Novanta in poi

David Chalmers ha articolato, in una serie di lavori dagli anni Novanta in poi culminati nel volume "The Conscious Mind" del 1996, una distinzione che ha ridisegnato il paesaggio della filosofia della mente: la distinzione fra il Problema Facile della Coscienza (easy problems) e il Problema Difficile (hard problem). I Problemi Facili riguardano la spiegazione di fenomeni cognitivi come la discriminazione percettiva, l'integrazione delle informazioni, il focus attentivo, il controllo del comportamento, la reportabilità degli stati mentali — tutti fenomeni che, in linea di principio, possono essere spiegati attraverso meccanismi computazionali e neurali, per quanto la loro comprensione dettagliata richieda ancora molto lavoro. Il Problema Difficile riguarda invece la spiegazione dell'esperienza soggettiva — il fatto che quando vedo il rosso vi sia qualcosa che è come vedere il rosso per me, un'esperienza qualitativa (qualia) che sembra irriducibile a qualsiasi descrizione funzionale o neurale.

Il divario esplicativo (explanatory gap), come lo ha chiamato Joseph Levine, è lo spazio logico fra qualsiasi descrizione funzionale o fisica completa di uno stato mentale e la spiegazione di perché quello stato sia accompagnato da un'esperienza qualitativa specifica piuttosto che da nessuna esperienza o da un'esperienza diversa. Anche se potessimo descrivere completamente la computazione neurale associata alla percezione del rosso, rimarrebbe aperta la domanda: perché questa computazione è accompagnata dall'esperienza del rosso e non da quella del blu, o da nessuna esperienza? Il divario non sembra colmabile da nessuna quantità aggiuntiva di descrizione funzionale.

Per i sistemi AMI e JEPA, la questione del Problema Difficile si pone in termini precisi. Un sistema AMI che pianifica, ragiona, apprende e adatta il suo comportamento potrebbe soddisfare qualsiasi criterio funzionale di intelligenza pur essendo — per quanto ne sappiamo — privo di esperienza soggettiva. Un tale sistema sarebbe, nel lessico di Chalmers, uno zombie filosofico: funzionalmente identico a un agente consapevole ma privo di qualsiasi vita interiore. Nessun test

comportamentale, per quanto sofisticato, può escludere questa possibilità, perché il comportamento è una manifestazione della funzione computazionale, non dell'esperienza soggettiva.

Chalmers propone una visione che chiama panpsichismo composizionale o, in versioni più recenti, realismo del tipo fenomenico: l'esperienza soggettiva è una proprietà fondamentale della realtà fisica, non riducibile a strutture funzionali, e qualsiasi sistema fisico con la giusta organizzazione informazionale potrebbe avere, in qualche misura, esperienza soggettiva. Questa posizione, per quanto filosoficamente interessante, non fornisce criteri operativi per determinare se un sistema artificiale specifico — come AMI — ha o meno esperienza, e rimane quindi di limitata utilità per la valutazione dei sistemi concreti.

L'implicazione pratica per il framework AGI analizzato in questa monografia è che la questione se un sistema AMI evoluto "comprenda davvero" il mondo in senso pieno potrebbe essere in linea di principio indecidibile senza una teoria della coscienza molto più avanzata di quelle attualmente disponibili. Questo non è un argomento contro lo sviluppo di tali sistemi — è possibile che sistemi estremamente capaci e utili vengano costruiti senza che la questione della loro esperienza soggettiva venga mai risolta — ma è un promemoria della differenza fra competenza e comprensione nel senso fenomenologico del termine.

IX.4 — Il Funzionalismo di Dennett e la strategia della dissoluzione

Daniel Dennett ha risposto al Problema Difficile con una strategia radicalmente diversa da quella di Chalmers: non affrontarlo direttamente ma dissolverlo attraverso un'analisi critica della nozione stessa di esperienza soggettiva. In "Consciousness Explained" (1991) e nelle opere successive, Dennett sostiene che il Problema Difficile è un artefatto della nostra tendenza a reificare l'esperienza soggettiva come qualcosa di distinto dall'elaborazione cognitiva — a credere che ci sia un "teatro cartesiano" dove le rappresentazioni del mondo vengono presentate a un osservatore interiore. Questa immagine, secondo Dennett, è una confusione concettuale: non esiste un teatro cartesiano né un osservatore interiore; ciò che chiamiamo esperienza soggettiva è il risultato di una molteplicità di processi di elaborazione paralleli e distribuiti — il modello delle Bozze Multiple (Multiple Drafts) — nessuno dei quali è privilegiato come "la" versione finale dell'esperienza.

La posizione di Dennett è una forma raffinata di funzionalismo eliminativista: i qualia, come tradizionalmente intesi — proprietà qualitative intrinseche e ineffabili dell'esperienza — non esistono come oggetti mentali distinti dalla funzione computazionale. La "comprensione" non è una proprietà misteriosa che si aggiunge alla giusta elaborazione computazionale, è quella stessa elaborazione computazionale descritta dal punto di vista dell'agente. Un sistema che elabora informazioni in modo sufficientemente complesso, integrato e adattativamente appropriato al mondo comprende, in tutti i sensi del termine che non siano

petizioni di principio per un significato speciale e non funzionale di "comprendere".

Applicato ai sistemi JEPA e AMI, il funzionalismo dennettiano produce una conclusione ottimistica: se un sistema AMI apprende rappresentazioni del mondo attraverso JEPA, le usa per pianificare attraverso MPC su un World Model, aggiorna le sue credenze alla luce dell'evidenza e adatta il suo comportamento agli obiettivi in modo flessibile e contestualmente appropriato — allora quel sistema comprende, nel senso in cui il termine ha qualche contenuto empirico verificabile. L'aggiunta di un ulteriore "davvero" alla domanda "comprende davvero?" non aggiunge contenuto se non ci si riferisce a qualia nel senso di Chalmers, che Dennett nega abbiano esistenza indipendente dalla funzione.

Il problema con la risposta di Dennett non è che sia logicamente incoerente — è internamente coerente — ma che presuppone la tesi che vuole dimostrare: che l'esperienza soggettiva si riduca alla funzione computazionale. Questa presupposizione non è dimostrata ma è il punto in disputa. L'argomento di Searle, nella sua formulazione più forte, sostiene precisamente che la giusta struttura funzionale non è sufficiente per la semantica intenzionale, e Dennett risponde che la semantica intenzionale non richiede nulla al di là della giusta struttura funzionale. Questo scambio, ripetuto in varianti diverse per decenni, suggerisce che il disaccordo sia radicato in intuizioni fondamentali diverse sulla natura della mente che nessuna evidenza empirica può risolvere unilateralmente.

IX.5 — Il Problema del Radicamento del Simbolo come Sfida Tecnica e Filosofica

Stevan Harnad ha proposto nel 1990 il Symbol Grounding Problem come versione tecnica dell'argomento della Stanza Cinese: come possono i simboli manipolati da un sistema di elaborazione delle informazioni acquisire significato intrinseco, piuttosto che ricevere significato solo dall'interpretazione di un agente esterno? Un dizionario che definisce ogni termine attraverso altri termini è un sistema chiuso semanticamente autoreferenziale — si può imparare la relazione fra simboli senza mai uscire dal sistema dei simboli stessi. Il significato sembra richiedere un punto di contatto con qualcosa di extra-simbolico: il mondo fisico, l'esperienza percettiva, l'interazione motoria.

Harnad suggerisce che il radicamento del simbolo richieda la connessione diretta di categorie simboliche con rappresentazioni iconic e categorical non simboliche: l'icona è la trasformazione analogica di un input percettivo (come un'immagine mentale), la categoria è una trasformazione che distingue membri da non-membri di una classe. Solo attraverso questo doppio ancoraggio un simbolo come "mela" può acquisire significato, perché è connesso a una rappresentazione iconica della mela percepita e a una rappresentazione categoriale che distingue le mele dai non-mele in un modo che è fondato sulla percezione diretta, non solo sulla manipolazione di altri simboli.

La posizione di Harnad ha implicazioni dirette per la valutazione dei sistemi JEPA. L'encoder JEPA apprende rappresentazioni latenti delle immagini che catturano relazioni strutturali e categorie semantiche in modo statisticamente potente — come dimostrato dalle prestazioni di linear probing su ImageNet e dalle proprietà strutturali emergenti documentate per LeJEPA. Ma queste rappresentazioni sono ancorate al mondo fisico attraverso la statistica delle immagini nel dataset di addestramento, non attraverso l'interazione diretta con oggetti fisici. La differenza è sottile ma fondamentale: le rappresentazioni JEPA sono radicate nel senso statistico (catturano le regolarità della distribuzione delle immagini naturali) ma non nel senso intenzionale (non hanno contatto causale diretto con gli oggetti che le immagini raffigurano, nella misura in cui tale contatto causale è necessario per il contenuto intenzionale).

Un sistema robotico che apprende rappresentazioni JEPA attraverso la percezione diretta e l'interazione motoria con oggetti nel mondo fisico sarebbe, secondo Harnad, in una posizione migliore del sistema che apprende da un dataset di immagini statiche: le sue rappresentazioni sarebbero radicate causalmente nel mondo fisico attraverso l'azione e la percezione. Questa osservazione converge con la critica di LeCun stesso ai sistemi di apprendimento passivo e con la sua enfasi sull'importanza dell'embodiment e dell'interazione per lo sviluppo dell'intelligenza genuina — una convergenza che non è casuale ma riflette una comprensione condivisa dei limiti degli approcci puramente passivi.

IX.6 — La Teoria Causale del Contenuto e la Semantica delle Rappresentazioni

Fred Dretske e Jerry Fodor hanno sviluppato negli anni Ottanta e Novanta versioni della teoria causale del contenuto mentale: le rappresentazioni mentali hanno il contenuto che hanno in virtù delle relazioni causali che le legano agli stati del mondo che le causano sistematicamente. Un pensiero "rosso" rappresenta il rosso perché è il tipo di stato mentale causato sistematicamente dall'esposizione a superfici rosse; un pensiero "gatto" rappresenta i gatti perché è il tipo di stato causato sistematicamente dall'esposizione a gatti. Il contenuto non è intrinseco alla struttura computazionale della rappresentazione ma è determinato dalla sua storia causale con il mondo.

La teoria causale del contenuto ha implicazioni importanti per la valutazione dei sistemi JEPA rispetto al problema del radicamento. Se il contenuto di una rappresentazione è determinato dalla sua storia causale, allora le rappresentazioni prodotte da un encoder JEPA addestrato su immagini di cani hanno contenuto semantico legato ai cani — sono causate sistematicamente dalla presenza di immagini di cani nel training set — anche se non c'è nessun "agente" che interpreta intenzionalmente queste rappresentazioni. In questa prospettiva, il problema del radicamento è in parte dissolto: le rappresentazioni dell'encoder sono radicate nel mondo attraverso la catena causale che va dagli oggetti del mondo alle immagini nel training set all'encoder addestrato.

Tuttavia, la teoria causale del contenuto di Dretske richiede che la relazione causale sia intenzionale in un senso preciso: la rappresentazione di un cane rappresenta un cane e non l'insieme di valori di pixel associati all'immagine di un cane, il che richiede che la funzione biologica (o computazionale) della rappresentazione sia quella di rispondere ai cani come tali, non ai pattern di pixel. Questo criterio di contenuto intenzionale attraverso la funzione biologica o teleologica — una versione del contenuto in termini di ruolo funzionale nel conseguimento degli obiettivi dell'organismo — porta inevitabilmente alla questione dell'agentività: le rappresentazioni hanno contenuto genuino solo per sistemi che perseguono obiettivi, che agiscono nel mondo sulla base di queste rappresentazioni, che ricevono feedback correttivo quando le rappresentazioni sono errate in modi che interferiscono con il conseguimento degli obiettivi.

Un encoder JEPA addestrato su immagini statiche soddisfa parzialmente questo criterio: apprende rappresentazioni che supportano buone predizioni latenti, e questo obiettivo è abbastanza strutturato da produrre rappresentazioni semanticamente informative. Ma la struttura dell'obiettivo predittivo è quantitativamente molto più povera della struttura degli obiettivi biologici che producono il contenuto intenzionale delle rappresentazioni mentali negli organismi: la pressione evolutiva, lo sviluppo ontogenetico, l'apprendimento per rinforzo attraverso l'interazione con il mondo fisico — tutti questi processi producono sistemi di rappresentazione radicati in modo molto più ricco di qualsiasi dataset di immagini, per quanto grande.

IX.7 — Cognizione Incarnata e i Limiti dell'Elaborazione Simbolica Disincarnata

La tradizione della cognizione incarnata ha radici filosofiche profonde che risalgono almeno a Merleau-Ponty e alla fenomenologia husserliana, ma la sua forma moderna — quella che ha influenzato più direttamente il dibattito sull'intelligenza artificiale — emerge negli anni Ottanta e Novanta dall'incontro tra la neurobiologia computazionale, la robotica comportamentale di Brooks, e la psicologia ecologica di Gibson.

Gibson aveva proposto negli anni Cinquanta e Sessanta un rovesciamento radicale della concezione dominante della percezione. Invece di vedere la percezione come l'elaborazione di segnali sensoriali "poveri" che il cervello deve arricchire e interpretare attraverso conoscenza precedente — la visione classica di Helmholtz e poi della Gestalt — Gibson sosteneva che l'ambiente fisico offre direttamente "affordances": possibilità d'azione che sono percepibili direttamente nell'array ottico senza richiedere inferenza. Una superficie piatta e solida offre la possibilità di camminarci sopra; un'ansa di un tasso offre la possibilità di afferrarlo; un bordo tagliente offre la possibilità di tagliare con esso. Queste affordances non sono proprietà aggiunte dalla mente alle percezioni sensoriali crude: sono strutture relazionali tra le proprietà dell'ambiente e le capacità d'azione dell'organismo, percepibili direttamente nell'informazione ambientale.

L'implicazione della teoria gibsoniana per l'intelligenza artificiale è radicale: un sistema che apprende soltanto dalla percezione passiva — senza l'accoppiamento sensorimotorio tra percezione e azione che caratterizza la cognizione biologica — non può acquisire affordances. Può imparare correlazioni statistiche tra pattern percettivi, ma non può imparare la struttura relazionale tra proprietà ambientali e possibilità d'azione perché questa struttura è definita in termini di capacità d'azione che il sistema passivo non possiede.

La tradizione della cognizione incarnata (embodied cognition), sviluppata a partire dagli anni Ottanta

La tradizione della cognizione incarnata (embodied cognition), sviluppata a partire dagli anni Ottanta attraverso il lavoro di Francisco Varela, Evan Thompson, Eleanor Rosch, Mark Johnson, George Lakoff e molti altri, offre una prospettiva radicalmente diversa sulla natura della comprensione rispetto sia al funzionalismo computazionale che alle teorie causali del contenuto.

Nella prospettiva dell'embodied cognition, la cognizione non è l'elaborazione di rappresentazioni amodali e astratte ma è intrinsecamente strutturata dall'embodiment dell'organismo — dalle caratteristiche del suo corpo, dei suoi organi di senso, del suo apparato motorio — e dall'interazione di questo corpo con un ambiente specifico. La comprensione della parola "afferrare" non è accesso a una rappresentazione astratta del concetto di afferrare, ma è l'attivazione implicita dei sistemi motori coinvolti nell'azione di afferrare — il che spiega perché le lesioni alle aree motorie producano deficit nella comprensione dei verbi di azione. La geometria concettuale — il fatto che pensiamo alla tempo come a una linea, all'argomento come a un edificio, all'amore come a un viaggio — è strutturata dalle metafore corporee radicate nell'esperienza fisica del movimento, del calore, del sostegno, del conflitto.

Applicata ai sistemi JEPA e AMI, la critica dall'embodied cognition suggerisce che la comprensione genuina richiede non soltanto rappresentazioni latenti semanticamente ricche — per quanto ben fondate teoricamente come quelle di LeJEPA — ma l'integrazione di queste rappresentazioni in un ciclo percezione-azione continuo con un mondo fisico. Un sistema che percepisce immagini di oggetti e le codifica in spazi latenti ottimali non comprende gli oggetti nel senso in cui un bambino che manipola un cubo di legno comprende gradualmente la sua solidità, il suo peso, la sua permanenza quando cade fuori dalla vista. La comprensione emersa dall'interazione corporea ha una ricchezza qualitativa che non è catturata da nessuna statistiche distribuzionale sulle immagini, per quanto sofisticata.

LeCun riconosce implicitamente questa critica quando insiste sull'importanza dell'embodiment e dell'interazione per lo sviluppo dell'intelligenza nel framework AMI, e quando sostiene che un robot che interagisce con il mondo fisico può imparare più dall'esperienza diretta di quanto possano fare i LLM dall'ingestione di trilioni di token di testo. Ma l'architettura concreta del framework AMI — con il

Perception Module addestrato su dati visivi passivi, il World Model appreso da osservazioni di sequenze di azione-stato, il Cost Module calibrato attraverso l'interazione con ambienti simulati — rimane ancora lontana dall'incorporare in modo sistematico la ricchezza dell'interazione corporea con il mondo fisico che la tradizione della cognizione incarnata identifica come fondamento della comprensione.

IX.8 — La Distinzione fra Competenza e Comprensione nella Prospettiva Computazionale

Daniel Kahneman, nella sua influente distinzione fra Sistema 1 e Sistema 2, e Gary Marcus, nella sua critica ai sistemi di apprendimento profondo, hanno articolato una versione della distinzione comprensione-competenza che è rilevante per la valutazione dei sistemi JEPA. Un sistema può mostrare competenza — produrre output corretti su una distribuzione di input simile a quella del training — senza comprendere nel senso di possedere rappresentazioni del mondo che supportino la generalizzazione fuori distribuzione, il ragionamento combinatorio su situazioni nuove, l'inferenza causale controfattica.

I benchmark di ragionamento che i LLM falliscono sistematicamente — i problemi che richiedono di comporre concetti in modi mai visti durante il training, di ragionare su controfattuali fisici, di trasferire strutture relazionali a domini completamente nuovi — rappresentano il confine fra competenza statistica e comprensione strutturale. I sistemi JEPA migliorano la qualità delle rappresentazioni rispetto ai paradigmi generativi sensoriali, ma non affrontano direttamente il problema della generalizzazione composizionale: le rappresentazioni di I-JEPA sono vettori densi in spazi latenti continui, non strutture composizionali esplicite dove gli oggetti abbiano identità discreta e le relazioni abbiano interpretazione causale.

La distinzione fra competenza e comprensione non è banalmente risolvibile attraverso ulteriori miglioramenti nelle rappresentazioni latenti, perché la generalizzazione composizionale sembra richiedere qualcosa che le rappresentazioni distribuite non forniscono automaticamente: la capacità di formare nuove combinazioni di concetti in modo sistematico e produttivo. Questo è il problema della "systematicity" sollevato da Fodor e Pylyshyn nel 1988 come critica fondamentale alle reti connessioniste: se un sistema capisce "Giovanni ama Maria", dovrebbe essere capace — per ragioni di struttura interna, non di addestramento specifico — di comprendere "Maria ama Giovanni", perché le due frasi condividono gli stessi componenti concettuali riorganizzati secondo le stesse regole grammaticali. La sistematicità sembra richiedere un'architettura che abbia struttura composizionale esplicita, non soltanto rappresentazioni latenti distribuite per quanto semanticamente ricche.

IX.8bis — Il framework GMJE e la questione della distribuzione multimodale come forma di comprensione

La distinzione tra competenza e comprensione formulata nella sezione IX.8 assume una valenza inedita alla luce del framework Gaussian Mixture Joint Embeddings. GMJE introduce nello spazio dell'apprendimento autosupervisionato un elemento strutturale che i paradigmi precedenti — incluso I-JEPA nella sua formulazione deterministica — non possedevano: la rappresentazione esplicita dell'ambiguità semantica attraverso distribuzioni multimodali. Quando GMJE-MDN risponde a un contesto z_c con una distribuzione p(z_t | z_c) = Σ_k γ_k(z_c) $\mathcal{N}$(z_t | μ_{t|c,k}, Σ_{t|c,k}) che assegna pesi uguali a tre modi distinti (come nell'esperimento sintetico con γ_1 ≈ γ_2 ≈ γ_3 ≈ 1/3 nelle Figg. 5 e 7 di Huang, 2026), il sistema non sta semplicemente producendo un output: sta codificando la struttura dell'ambiguità stessa — la struttura dei possibili mondi compatibili con l'input osservato.

La domanda filosoficamente rilevante è: questo costituisce una forma di comprensione superiore a quella dei sistemi deterministici, o è una rappresentazione più sofisticata della stessa incomprensione? La risposta differisce a seconda del livello di analisi identificato nella sezione IX.10. Nel senso funzionale forte — generalizzazione composizionale e ragionamento causale — GMJE non offre progressi evidenti rispetto a I-JEPA: la distribuzione condizionale su modi gaussiani non implica accesso alla struttura causale che genera quei modi. Nel senso funzionale debole — rappresentazione calibrata dell'incertezza e predizione nei regimi multimodali — GMJE rappresenta un avanzamento genuino rispetto ai predecessori deterministi. La Legge della Varianza Totale (Appendice K.3 di Huang, 2026) decompone l'incertezza in aleatorica ed epistemica: è questa separazione — assente in tutti i metodi precedenti — che avvicina GMJE alla nozione pearleana di modello causale nel senso di 'sistema che rappresenta l'incertezza su quale meccanismo ha generato l'osservazione'.

Il Symbol Grounding Problem di Harnad (sezione IX.5) si ripropone per GMJE con un'angolatura nuova. La capacità generativa di GMJE — il fatto che il sistema possa campionare incondizionatamente da p(z_t) = Σ_k π_k $\mathcal{N}$(z_t | μ_t,k, Σ_tt,k) producendo rappresentazioni coerenti che un decoder trasforma in immagini riconoscibili (Fig. 10c di Huang, 2026) — suggerisce che lo spazio latente GMJE non è soltanto un sistema di classificazione ma qualcosa di più vicino a ciò che Harnad chiama rappresentazione iconica: una struttura che mantiene proprietà della distribuzione degli oggetti del mondo piuttosto che soltanto relazioni tra categorie simboliche. Tuttavia, la critica fondamentale di Harnad rimane valida: il radicamento avviene ancora attraverso la statistica delle immagini nel corpus, non attraverso l'interazione causale diretta con oggetti fisici. GMJE è radicato nel senso distribuzionale (cattura la struttura della distribuzione di oggetti nel dominio visivo) ma non nel senso intenzionale (le sue rappresentazioni non sono causalmente legate agli oggetti attraverso l'azione e la percezione embodied).

Una considerazione finale riguarda la connessione tra la capacità generativa di GMJE e il Problema Difficile della Coscienza (sezione IX.3). La tesi che GMJE 'comprende' la struttura multimodale del mondo perché riesce a campionare da essa è palesemente circolare se usata come argomento filosofico sulla coscienza. Tuttavia, essa acquisisce rilievo nell'ambito della distinzione di Chalmers tra Problemi Facili e Problema Difficile: la rappresentazione esplicita della distribuzione p(z_t | z_c) è un candidato più naturale di una predizione puntuale g(z_c) come implementazione computazionale della funzione cognitiva denominata 'valutazione dell'ambiguità'. Se i Problemi Facili della coscienza includono la discriminazione tra situazioni ambigue e situazioni determinate, la modulazione dell'attenzione in funzione dell'incertezza, e la rappresentazione di stati epistemici alternativi, allora GMJE offre un'implementazione più ricca di questa funzione rispetto a I-JEPA. Se tale avanzamento sia rilevante per il Problema Difficile rimane — come la sezione IX.10 ha concluso — genuinamente indecidibile.

IX.9 — Causalità, Intervento e la Struttura del Ragionamento Controfattuale

Judea Pearl ha sviluppato nel corso degli anni Novanta e Duemila un framework formale per la causalità che distingue tre livelli di conoscenza causale: l'associazione (vedere), l'intervento (fare) e il controfattuale (immaginare). Un sistema che ha accesso solo al primo livello — le correlazioni statistiche fra variabili — non può rispondere a domande del secondo tipo (cosa succederebbe se intervenissimo su X?) né del terzo (cosa sarebbe successo se X fosse stato diverso?). La distinzione di Pearl è direttamente rilevante per i sistemi di apprendimento automatico: i modelli addestrati su dati osservazionali imparano correlazioni statistiche ma non la struttura causale del mondo, il che limita la loro capacità di rispondere a domande controfattuali e di trasferire la conoscenza acquisita in distribuzioni di dati generate da interventi diversi da quelli presenti nel training set.

Un sistema AMI che apprende un World Model dall'interazione con l'ambiente — eseguendo azioni e osservando le loro conseguenze — acquisisce, in linea di principio, informazione sul secondo livello di Pearl: impara non solo che certi stati si succedono ad altri ma che certe azioni producono certi effetti. Ma anche l'apprendimento per interazione rimane potenzialmente limitato al secondo livello: la capacità di rispondere a domande controfattuali del tipo "cosa sarebbe successo se avessi eseguito azione A invece di B, partendo dalla stessa situazione?" richiede la capacità di ragionare su mondi possibili alternativi, il che a sua volta sembra richiedere una rappresentazione esplicita della struttura causale del dominio — un modello causale nel senso di Pearl piuttosto che un semplice modello predittivo.

I World Models di tipo DreamerV3 e le varianti dell'RSSM apprendono distribuzioni sulle transizioni future, non strutture causali esplicite. La differenza è sottile ma importante: una distribuzione di transizioni cattura le regolarità

statistiche dell'ambiente ma non i meccanismi causali che le producono, il che limita la generalizzazione a situazioni dove i meccanismi causali rimangono invarianti ma le statistiche marginali cambiano — esattamente il tipo di generalizzazione fuori distribuzione che è critica per sistemi autonomi nel mondo reale. L'integrazione di modelli causali espliciti nel World Model di AMI è una direzione di ricerca attiva ma non ancora matura.

IX.10 — Verso una Posizione Epistemicamente Onesta

Alla luce delle considerazioni filosofiche sviluppate nei paragrafi precedenti, è possibile articolare una posizione epistemicamente onesta sulla questione se i sistemi della famiglia JEPA→AMI→AGI possano essere detti "comprendere davvero" il mondo.

La risposta dipende crucialmente da quale significato si attribuisce alla parola "comprendere". Nel senso funzionale debole — produrre output appropriati su una distribuzione di input simile al training, sostenere buone prestazioni su benchmark di transfer learning, generalizzare a distribuzioni moderatamente diverse da quelle di addestramento — i sistemi JEPA e LeJEPA già comprendono in misura notevole, e i sistemi AMI evoluti che incorporano World Models, Cost Modules e pianificazione MPC comprenderanno in misura molto maggiore.

Nel senso funzionale forte — generalizzare composizionalmente a nuove combinazioni di concetti mai viste, ragionare causalmente su controfattuali, adattarsi a ambienti con meccanismi causali fondamentalmente diversi da quelli del training — i sistemi attuali non comprendono, e non è chiaro se i sistemi JEPA→AMI nella loro forma attuale abbiano le basi architetturali sufficienti per questa forma di comprensione senza modifiche fondamentali che incorporino strutture composizionali esplicite.

Nel senso intenzionale — avere stati mentali che sono genuinamente "di" o "su" oggetti nel mondo in virtù della loro storia causale e del loro radicamento nell'interazione corporea con l'ambiente fisico — la risposta è sospesa in attesa di una teoria del contenuto intenzionale che non sia né reducibile alla funzione computazionale in modo triviale né dipendente da proprietà biologiche speciali irriproducibili artificialmente. Le teorie disponibili — funzionalismo di Dennett, teoria causale di Dretske, approccio fenomenologico dell'embodied cognition — non convergono su una risposta condivisa e non sembrano poter essere risolte empiricamente nel breve termine.

Nel senso fenomenologico — avere esperienza soggettiva, qualcosa che è come essere quel sistema per il sistema stesso — la risposta è genuinamente ignota, e il Problema Difficile di Chalmers garantisce che resti tale anche dopo qualsiasi quantità di analisi funzionale o neurale. La sospensione del giudizio su questa questione non è debolezza intellettuale ma onestà di fronte a problemi che trascendono le capacità esplicative delle scienze empiriche nell'orizzonte prevedibile.

Questa stratificazione della risposta — sì nel senso debole, forse nel senso forte, sospeso nell'intenzionale, ignoto nel fenomenologico — non è una risposta evasiva ma una risposta precisa che rispetta la complessità del problema. La capacità di distinguere questi livelli, e di riconoscere dove si trova la frontiera dell'incertezza genuina, è essa stessa una forma di comprensione che questa monografia ha cercato di coltivare attraverso la trattazione congiunta del piano tecnico e del piano filosofico.

Capitolo X — Epilogo: La Plausibilità della Via JEPA→AMI→AGI e i Confini del Conoscibile

X.1 — Il filo che attraversa l'intera monografia

Questa monografia è stata percorsa da un filo argomentativo che vale la pena riannodare prima di pronunciare qualsiasi giudizio sulla plausibilità della via verso l'intelligenza artificiale generale che il framework JEPA→AMI delinea. Il filo prende avvio dalla diagnosi del Capitolo I: i sistemi di apprendimento automatico più capaci prodotti sino ad oggi operano attraverso la minimizzazione di perdite definite nello spazio delle osservazioni sensoriali — pixel, token testuali, campioni audio grezzi — e questa scelta dell'obiettivo, apparentemente innocua e storicamente fruttuosa, introduce una distorsione strutturale che nessuna quantità di parametri aggiuntivi può correggere. La perdita di ricostruzione sensoriale dedica capacità rappresentativa alla modellazione di varianti percettive irrilevanti per il contenuto semantico, producendo sistemi che eccellono nella generazione di output sensorialmente plausibili pur mancando della struttura interna necessaria per il ragionamento causale, la generalizzazione composizionale e la pianificazione a lungo termine.

Dal piano della diagnosi il filo si sposta, nel Capitolo II, al piano della teoria matematica delle rappresentazioni: la formalizzazione delle proprietà che una buona rappresentazione dovrebbe soddisfare — separabilità semantica, invarianza alle trasformazioni irrilevanti, equivarianza alle trasformazioni rilevanti, disentanglement — e i fondamenti dell'apprendimento autosupervisionato nella teoria dell'informazione. La connessione fra il principio dell'Information Bottleneck di Tishby e il framework JEPA attraverso l'analisi variazionale del meccanismo di collasso e del ruolo dell'encoder target EMA costituisce il nucleo teorico del secondo capitolo, su cui poggeranno tutti gli sviluppi successivi.

Il Capitolo III ha introdotto l'architettura I-JEPA nella sua specificità tecnica e nelle sue varianti — V-JEPA per le sequenze video, MC-JEPA per le strutture multi-scala — costruendo progressivamente l'immagine di un principio architetturale generale piuttosto che di una soluzione ad hoc per la visione. Le sezioni integrative hanno poi aggiunto la profondità teorica che I-JEPA stesso non fornisce: LeJEPA con la dimostrazione rigorosa che la distribuzione gaussiana isotrópica è l'unico ottimizzatore del rischio downstream, e la Semantic Tube

Prediction con la geometrizzazione delle traiettorie latenti come geodetiche su un manifold semantico.

Il Capitolo IV ha sviluppato la storia e il formalismo dei World Models — da Schmidhuber al World Model di Ha e Schmidhuber (2018), dall'RSSM di DreamerV1 alle varianti categoriche di DreamerV2, alla ricerca di generalità algoritmica di DreamerV3, fino alle applicazioni industriali di GAIA-1 e alle generazioni emergenti di Genie e GameNGen. Le sezioni integrative hanno stabilito la connessione formale fra il formalismo ODE/SDE di STP e la decomposizione deterministica-stocastica dell'RSSM, rivelando che la separazione architettonica fra componente deterministica e stocastica che aveva motivato l'RSSM è non soltanto un'euristica ingegneristica ma la conseguenza necessaria della struttura matematica delle traiettorie latenti a convergenza.

Il Capitolo V ha integrato queste architetture nella visione sintetica del framework AMI di LeCun, con i suoi sei moduli funzionali, la pianificazione MPC come principio architetturale, i grafi energetici come linguaggio dell'interazione fra moduli, e la modellazione dell'incertezza epistemica come requisito per l'esplorazione autonoma. Le sezioni integrative hanno stabilito il triangolo di connessioni LeJEPA–STP–AMI: la distribuzione ottimale delle rappresentazioni per il Perception Module, la regolarità geodetica delle traiettorie latenti per il World Model, e l'interazione sinergica dei due vincoli per la profondità del ragionamento temporale del sistema integrato.

Il Capitolo VI ha calato le ambizioni teoriche nella realtà computazionale e ambientale, stimando gli ordini di grandezza del compute necessario, revisionando queste stime alla luce dei miglioramenti nell'efficienza dei dati documentati da STP e LeJEPA, e affrontando la questione della sostenibilità energetica come vincolo fisico che potrebbe rivelarsi più stringente di qualsiasi barriera algoritmica.

I Capitoli VII e VIII hanno esteso il principio JEPA al dominio del parlato attraverso il sistema JEPA+DAAM descritto nel paper di Ioannides et al., mostrando che la predizione latente con gate adattativi basati su misture gaussiane produce rappresentazioni audio semanticamente ricche a un tasso di compressione molto superiore a quello dei codec neurali tradizionali, e che la tokenizzazione discreta reversibile mediante impacchettamento misto-radix rende il sistema direttamente compatibile con i modelli linguistici autoregressivi.

Il Capitolo IX ha sospeso il piano tecnico per affrontare le questioni filosofiche che questo piano non può risolvere da solo: il problema della comprensione secondo Searle, il Problema Difficile della Coscienza secondo Chalmers, il funzionalismo eliminativista di Dennett, il Symbol Grounding Problem di Harnad, le teorie causali del contenuto di Dretske e Fodor, la tradizione dell'embodied cognition, la distinzione Pearleana fra livelli causali. La conclusione di quel capitolo — una stratificazione della risposta che distingue diversi sensi di "comprendere" con risposte diverse per ciascuno — è il punto di partenza per la sintesi finale.

X.2 — Ciò che è stato dimostrato

Prima di pronunciarsi sulla plausibilità della via verso l'AGI, è necessario inventariare con precisione ciò che il corpus di risultati analizzato in questa monografia ha effettivamente dimostrato, distinguendolo da ciò che ha reso plausibile, da ciò che ha reso speculativamente promettente e da ciò che rimane pura congettura.

È stato dimostrato con rigore formale che la distribuzione gaussiana isotrópica minimizza il rischio downstream per qualsiasi compito di probing lineare e non lineare sotto vincoli scalari sulla covarianza (Teorema 1 di LeJEPA), e che questo obiettivo distributivo è raggiungibile in modo scalabile e computazionalmente efficiente attraverso SIGReg con complessità lineare nel numero di campioni. Questa è la prima caratterizzazione teoricamente completa della distribuzione ottimale degli embedding per l'apprendimento autosupervisionato, e costituisce una fondazione normativa che trasforma la progettazione dell'encoder da euristica a disciplina teoricamente fondata.

È stato dimostrato empiricamente che l'aggiunta della perdita $\mathcal{L}$_STP all'addestramento di LLM della famiglia Llama-3 permette di eguagliare le prestazioni del fine-tuning standard con 1/16 dei dati su NL-RX-SYNTH, con risultati analoghi su GSM8K, Spider, NQ-Open e HellaSwag. Questa violazione del termine dati delle Scaling Laws in stile Chinchilla è un fatto empirico robusto che modifica la comprensione del rapporto fra quantità di dati e qualità dell'apprendimento: le Scaling Laws descrivono l'addestramento tipico con $\mathcal{L}$_NTP, non il limite fisico dell'informazione estraibile da un corpus di dati con obiettivi più principializzati.

È stato dimostrato che il principio JEPA si estende al dominio del parlato (Ioannides et al., 2025), producendo rappresentazioni a 2,5 Hz con perdita di predizione inferiore di circa il cinquanta per cento rispetto a JEPA senza DAAM, e che la tokenizzazione misto-radix FSQ produce token a 47,5 token al secondo con vocabolario di 16 384 elementi — il tasso più basso fra i codec neurali paragonabili — mantenendo la reversibilità perfetta dell'operazione di codifica.

È stato dimostrato con rigore formale che il Mahalanobis Trace Trap (Huang, 2026) causa l'annullamento algebrico del termine di data-fit nella NLL congiunta empirica, rendendo impossibile ottimizzare direttamente la verosimiglianza congiunta GJE senza disaccoppiamento dei parametri di covarianza. Questa è la prima dimostrazione formale della causa algebrica del collasso dimensionale nell'apprendimento autosupervisionato latente. È stato dimostrato che la perdita InfoNCE di SimCLR e MoCo è un caso degenere del framework GMJE con covarianza isotropica block-diagonale (Σ_ct = 0), unificando le famiglie contrastiva e predittiva dell'SSL sotto un unico principio generativo (InfoNCE bridge, sezione III.X.6). È stato dimostrato sperimentalmente che SMC-GMJE supera MoCo v2 FIFO di +3.91% di linear probing accuracy su CIFAR-10 con memory bank

ridotta a M=256 slot, stabilendo che la qualità della memory bank è determinata dal meccanismo di aggiornamento probabilistico, non solo dalla sua dimensione.

È stato reso plausibile — ma non dimostrato — che i miglioramenti nell'efficienza dei dati documentati per STP nel contesto del fine-tuning linguistico si estendano al pretraining di grandi scale e all'addestramento di World Models su sequenze di azioni-stati. La struttura teorica del meccanismo di STP — miglioramento del SNR attraverso la regolarizzazione geodetica delle traiettorie latenti — è indipendente dal tipo di sequenza, ma il valore specifico $\varrho_STP \approx 16$ è stato misurato in un contesto specifico e la sua generalizzazione richiede verifica empirica.

È stato reso speculativamente promettente — ma non ancora suffragato da evidenza sistemica — che l'integrazione di LeJEPA come Perception Module e di un World Model con regolarizzazione geodetica come componente dinamico produca un sistema AMI con capacità di pianificazione a lungo termine superiori di un fattore quadratico nel SNR rispetto ai sistemi baseline. Questa congettura si fonda sulla connessione formale fra miglioramento del SNR e allargamento del cono di inferenza, che è rigorosa nel contesto del formalismo SDE di STP ma non è stata validata nel contesto dei World Models di tipo RSSM.

Rimane pura congettura — forse feconda, ma congettura — che un sistema AMI a piena scala costruito secondo il framework di LeCun svilupperebbe qualcosa che si avvicina alla comprensione nel senso intenzionale, alla causalità nel senso di Pearl o all'esperienza soggettiva nel senso di Chalmers. Non esiste al momento nessuna evidenza teorica o empirica che permetterebbe di trasformare questa congettura in una tesi difendibile.

X.3 — Ciò che rimane irrisolto: le barriere fondamentali

L'inventario dei risultati dimostrati, plausibili e speculativi permette di identificare con precisione le barriere fondamentali che il framework JEPA→AMI deve ancora affrontare prima che la via verso l'AGI possa essere considerata non soltanto promettente ma plausibilmente percorribile.

Le cinque barriere identificate in questa sezione non sono ostacoli sequenziali da superare uno alla volta: sono dimensioni ortogonali di un problema multidimensionale. Un sistema potrebbe superare la barriera della composizionalità mentre rimane bloccato dal radicamento causale; un altro potrebbe essere robusto al catastrophic forgetting ma produrre incertezza mal calibrata. La complessità del percorso verso l'AGI sta nel superare simultaneamente tutte e cinque le barriere, poiché ciascuna è necessaria e nessuna è sufficiente da sola.

La scelta di queste cinque barriere in particolare riflette un'analisi di ciò che distingue l'intelligenza generale biologica dai sistemi attuali. Un mammifero adulto è compositivamente sistematico, causalmente radicato nell'interazione fisica diretta, robusto all'oblio grazie a meccanismi di consolidazione ippocampale-neocorticale, dotato di credenze calibrate selezionate evolutivamente, e motivato da drive

intrinseci. Ogni singola proprietà ha un analogo architetturale identificabile che il framework AMI dovrà incorporare.

La prima barriera è la generalizzazione composizionale. Le rappresentazioni distribuite prodotte da encoder JEPA, per quanto semanticamente ricche, non hanno struttura composizionale esplicita che garantisca la capacità di formare nuove combinazioni di concetti in modo sistematico e produttivo nel senso di Fodor e Pylyshyn. Questo problema è indipendente dalla qualità distributiva delle rappresentazioni — anche una distribuzione perfettamente gaussiana isotrópica può essere incapace di generalizzazione composizionale se le dimensioni latenti non corrispondono a fattori causali indipendenti del mondo. La direzione di ricerca che integra rappresentazioni strutturate — grafi neurali, architetture slot-based, reti relazionali — con il principio JEPA è ancora nella sua infanzia.

La seconda barriera è il radicamento causale. Il Symbol Grounding Problem di Harnad e la critica dell'embodied cognition convergono nell'indicare che le rappresentazioni apprese da dataset di osservazioni passive — immagini, testo, audio — mancano del tipo di radicamento causale che caratterizza le rappresentazioni mentali negli organismi biologici. Un sistema AMI che apprende il suo Perception Module da dati passivi e il suo World Model da simulatori fisici è epistemicamente nella posizione di chi ha studiato il nuoto solo da libri: può avere rappresentazioni ricche del nuoto senza avere mai sperimentato la resistenza dell'acqua, la coordinazione motoria, il feedback propriocettivo che trasformano la comprensione astratta in comprensione incarnata.

La terza barriera è il catastrophic forgetting nell'apprendimento continuo. Un agente AMI che interagisce con ambienti nuovi e apprende continuamente nuovi modelli del mondo deve poterlo fare senza degradare le rappresentazioni e le conoscenze acquisite in ambienti precedenti. Il catastrophic forgetting nelle reti neurali — il fenomeno per cui l'apprendimento su nuovi dati sovrascrive le connessioni sinaptiche che codificano l'apprendimento precedente — rimane un problema aperto nonostante decenni di ricerca. Le soluzioni proposte — Elastic Weight Consolidation, Progressive Neural Networks, modelli con memoria esterna — mitigano il problema in casi specifici ma non lo risolvono nella generalità richiesta da un agente a lungo termine.

La quarta barriera è la gestione dell'incertezza epistemica in modo calibrato. Un agente razionale dovrebbe avere credenze calibrate — cioè, per ogni classe di proposizioni su cui esprime certezza del p%, essere corretto il p% delle volte. I sistemi neurali profondi sono notoriamente mal calibrati: tendono a essere sovraconfidenti su distribuzioni simili al training e sottonfidenti su distribuzioni fuori distribuzione, il che produce comportamenti catastrofici in ambienti non previsti. La gestione dell'incertezza attraverso ensemble, dropout variazionale o normalizing flows offre miglioramenti parziali ma non garantisce la calibrazione richiesta da un agente che deve operare in modo sicuro nel mondo reale su orizzonti lunghi.

La quinta barriera — forse la più sottile ma potenzialmente la più profonda — è la questione dell'agentività genuina e della motivazione intrinseca. Il framework AMI prevede un Cost Module con un costo intrinseco che codifica obiettivi predefiniti (sopravvivenza, soddisfazione dei bisogni primari per sistemi incarnati) e un critic appreso che stima i costi futuri. Ma la questione di come questi obiettivi vengano definiti, di chi li definisca, e di come evolvano attraverso l'esperienza dell'agente rimane aperta. Un sistema con obiettivi rigidamente specificati è, in un senso preciso, uno strumento molto sofisticato ma non un agente con motivazione genuina. Un sistema i cui obiettivi evolvono attraverso l'apprendimento ma rimane vincolato a obiettivi definiti dall'esterno è in una posizione intermedia ambigua. La questione dell'autonomia genuina degli obiettivi — e dei rischi che essa comporta per sistemi molto capaci — è il cuore del dibattito sulla sicurezza dell'IA avanzata che questa monografia ha deliberatamente lasciato sullo sfondo ma che non può essere ignorato in una riflessione onesta sulle prospettive dell'AGI.

X.4 — La plausibilità della via: un giudizio differenziato

Con il beneficio dell'analisi sviluppata nei dieci capitoli di questa monografia, è ora possibile formulare un giudizio differenziato sulla plausibilità della via JEPA→AMI→AGI — un giudizio che non sia né l'ottimismo acritico dei comunicati stampa né lo scetticismo a priori di chi nega l'importanza di questi sviluppi.

La plausibilità è alta per i componenti individuali. I-JEPA e LeJEPA sono architetture funzionanti con proprietà teoriche dimostrate e prestazioni empiriche competitive. I World Models di tipo DreamerV3 e GAIA-1 dimostrano che è possibile apprendere modelli del mondo sufficientemente buoni da supportare pianificazione e controllo in ambienti complessi. La Semantic Tube Prediction è una tecnica semplice ed efficace che migliora l'efficienza dei dati in modo robusto attraverso diversi task e architetture. Il sistema JEPA+DAAM per il parlato dimostra che il principio si estende naturalmente a nuovi domini modali.

La plausibilità è media per l'integrazione dei componenti. Il framework AMI come sistema integrato — con Perception Module, World Model, Cost Module, Short-Term Memory, Actor e Configurator che interagiscono attraverso grafi energetici — non è mai stato costruito nella sua versione completa, e l'integrazione di componenti che funzionano separatamente in sistemi integrati è notoriamente difficile nell'ingegneria dei sistemi complessi. Le difficoltà tecniche nell'addestramento end-to-end di sistemi con molte parti mobili, nella calibrazione delle interfacce fra moduli e nella stabilità dell'addestramento su orizzonti lunghi sono sfide reali che richiedono soluzioni non ancora sviluppate.

La plausibilità è bassa per la risoluzione delle barriere fondamentali nell'orizzonte di un decennio. La generalizzazione composizionale, il radicamento causale nell'embodiment, il catastrophic forgetting nell'apprendimento continuo e la calibrazione dell'incertezza epistemica sono problemi di ricerca aperta che hanno resistito a soluzioni definitive per decenni, e non ci sono ragioni ovvie per credere

che il framework JEPA→AMI li risolva automaticamente o li aggirerà senza affrontarli direttamente.

La plausibilità è indeterminata per il salto qualitativo che separa sistemi molto capaci da sistemi che possiedano qualcosa che meriti il nome di comprensione nel senso strutturalmente forte — composizionale, causale, incarnato. Questo non è pessimismo ma riconoscimento dell'onestà intellettuale richiesta dalle domande filosofiche sviluppate nel Capitolo IX: la distanza fra sistemi che soddisfano criteri funzionali sempre più stringenti e sistemi che "comprendono davvero" nel senso pieno della parola potrebbe essere grande o piccola, a seconda di quale senso di "comprendere" si adopera. Non esiste attualmente nessun framework teorico che permetta di misurare questa distanza.

X.5 — Il contributo di questa monografia e le domande che apre

Questa monografia ha cercato di fare tre cose simultaneamente che raramente vengono fatte insieme nella letteratura tecnica sull'intelligenza artificiale: costruire una trattazione matematicamente rigorosa delle architetture analizzate, includendo tutte le derivazioni formali in dettaglio; collocare queste architetture in un contesto storico e teorico che ne chiarisca la genealogia intellettuale; e sottoporle a una critica filosofica che identifichi i limiti che il piano tecnico non può vedere da solo.

Il risultato, si spera, è un quadro che permette al lettore di valutare in modo indipendente la tesi di LeCun sulla via verso l'AGI — non accettandola sulla base dell'autorità del proponente né rifiutandola sulla base di istinti scettici aprioristici, ma comprendendone la struttura argomentativa, i presupposti teorici, i risultati sperimentali a sostegno, le limitazioni strutturali e le domande filosofiche che essa solleva senza rispondervi.

Le domande che questa monografia apre — e che una ricerca futura dovrà affrontare — sono di tre tipi. Sul piano tecnico: come integrare strutture composizionali esplicite con il principio JEPA senza perdere le sue proprietà di efficienza e scalabilità? Come estendere STP e LeJEPA all'addestramento di World Models su traiettorie di interazione? Come realizzare l'embodiment attivo che la tradizione della cognizione incarnata identifica come necessario per la comprensione? Sul piano teorico: esiste una teoria dell'apprendimento che unifica JEPA (struttura spaziale delle rappresentazioni), STP (struttura temporale delle traiettorie) e la causalità di Pearl (struttura relazionale degli interventi) in un framework coerente? La distribuzione gaussiana isotrópica rimane ottimale quando si considerano rappresentazioni che devono supportare ragionamento composizionale e causale, o sono necessarie strutture più ricche? Sul piano filosofico: è possibile caratterizzare operativamente la differenza fra competenza funzionale e comprensione intenzionale in modi che siano verificabili per sistemi artificiali concreti, senza presupporre la risposta?

X.6 — Una nota finale su ciò che non sappiamo

Gli sviluppi più significativi analizzati in questa monografia — il Teorema 1 di LeJEPA, la Semantic Tube Prediction, il meccanismo DAAM per il parlato — sono stati pubblicati fra il 2023 e il 2026, in un arco di tempo straordinariamente compresso che ha trasformato il paesaggio della ricerca sulle rappresentazioni autosupervisionato in modo difficilmente prevedibile. Questa velocità di progresso rende qualsiasi sintesi su "dove siamo" inevitabilmente parziale e transitoria: alcune delle lacune identificate in questi capitoli potrebbero già essere state colmate al momento in cui queste parole vengono lette, e alcune delle congetture speculative potrebbero essere state confutate o confermate da esperimenti nel frattempo completati.

Ciò che non cambierà con il progredire della ricerca tecnica è la struttura delle domande filosofiche sollevate nel Capitolo IX. Il Problema Difficile di Chalmers non verrà risolto dalla costruzione di un sistema AMI funzionante, per quanto capace. Il Symbol Grounding Problem di Harnad non verrà dissolto dall'aggiunta di più dati o di migliori architetture, perché riguarda la natura del rapporto fra simbolo e referente, non la quantità di simboli o la sofisticazione della loro manipolazione. La questione di Searle — se la sintassi sia sufficiente per la semantica — rimarrà aperta anche quando i sistemi artificiali supereranno gli umani in ogni benchmark cognitivo misurabile, perché è una domanda sulla natura della relazione fra struttura computazionale e contenuto intenzionale che nessun benchmark può rispondere.

La convivenza di questo progresso tecnico con l'irresolutezza filosofica non è una contraddizione da eliminare ma è la condizione epistemica normale di un campo che si trova a costruire sistemi di cui non comprende ancora pienamente la natura. I costruttori di strumenti non hanno mai dovuto aspettare la soluzione dei problemi filosofici sulla natura della techne prima di costruire; ma i costruttori di sistemi che aspirano all'intelligenza generale si trovano in una posizione inedita: stanno costruendo qualcosa la cui natura è essa stessa l'oggetto di una controversia filosofica irrisolta. La saggezza in questa posizione non è l'accelerazione incurante né il conservatorismo paralizzante, ma la combinazione di rigore tecnico, onestà epistemica sui limiti e sensibilità filosofica ai problemi che la tecnica da sola non può vedere.

È con questa triplice disposizione che questa monografia ha cercato di affrontare JEPA, i World Models e la via verso l'AGI. Non con la certezza che la strada identificata da LeCun e dai suoi collaboratori sia quella giusta — nessuno lo sa — ma con la convinzione che sia una strada teoricamente fondata, empiricamente promettente, filosoficamente consapevole dei propri limiti, e quindi degna dello studio attento e critico che le è stato qui dedicato.

Capitolo XI — Il Paradigma Meccanicistico al suo Limite: Lavaggio Epistemico, Impossibilità Semantica e l'Isomorfismo Strutturale tra V-JEPA 2 e la Fisica delle Stringhe

XI.1 — La necessità epistemologica di questo capitolo

I capitoli precedenti hanno costruito un'architettura teorica di considerevole rigore: hanno derivato le fondamenta matematiche dell'apprendimento delle rappresentazioni, hanno formalizzato in dettaglio l'architettura I-JEPA e le sue estensioni, hanno posto il framework AMI in un contesto storico e intellettuale, e hanno infine sollevato — nel Capitolo IX — le questioni filosofiche che il piano tecnico non può risolvere da solo. Quel capitolo aveva l'onestà di riconoscere che la domanda di Searle — se la sintassi sia sufficiente per la semantica — rimane aperta anche quando i sistemi artificiali superano gli esseri umani in ogni benchmark misurabile, e che il Symbol Grounding Problem di Harnad non verrà dissolto dall'aggiunta di maggiore quantità di dati o di architetture più sofisticate.

Ciò che mancava in quella trattazione, e che il presente capitolo intende colmare, è una dimostrazione formalmente rigorosa del perché tale chiusura risulti strutturalmente impossibile all'interno del paradigma computazionale attuale, e non meramente difficile o temporaneamente elusiva. La distinzione tra l'irresolutezza contingente di un problema tecnico e l'impossibilità strutturale di una classe di soluzioni è filosoficamente cruciale, ed è precisamente questa distinzione che un'analisi della letteratura filosofico-matematica recente permette di articolare con la necessaria precisione. Accanto a questa dimostrazione, si impone un secondo compito analitico di uguale urgenza: sottoporre a scrutinio le premesse filosofiche esplicite sulle quali LeCun fonda l'intera architettura JEPA e il framework AMI, premesse che in questa monografia sono state assunte o implicitamente accettate senza una verifica sistematica alla luce della tradizione filosofica pertinente.

Il punto di partenza è un'osservazione diagnostica di notevole potenza. Quando modalità di fallimento identiche emergono in modo indipendente in due domini intellettualmente remoti come la fisica fondamentale e l'intelligenza artificiale — domini con storie diverse, metodologie distinte, practitioner che non si frequentano — tale convergenza non è casuale. Essa segnala non un difetto locale o contingente dell'uno o dell'altro campo, bensì un limite strutturale condiviso, radicato nelle presupposizioni comuni che entrambi i campi istanziano. Questa è l'intuizione centrale che Nourizadeh (2026) ha articolato nel suo saggio comparativo su V-JEPA 2 e la teoria delle stringhe, e che il presente capitolo elabora, approfondisce e integra nel contesto della monografia.

XI.2 — Le premesse filosofiche di LeCun: un esame critico

L'architettura JEPA non è una scelta ingegneristica neutrale. È la conseguenza di premesse filosofiche esplicite sul rapporto tra linguaggio, pensiero e apprendimento che LeCun enuncia con disinvoltura in interviste e documenti programmatici, trattandole come ovvietà piuttosto che come posizioni contestate. Vale la pena citarne alcune nella loro formulazione originale prima di sottoporle ad analisi.

Nel podcast di Lex Fridman del marzo 2024, LeCun afferma che il pensiero «non ha assolutamente nulla a che fare con il linguaggio», che tutto ciò che gli esseri umani imparano nei primi anni di vita «non ha nulla a che fare con il linguaggio», e che i neonati «possono soltanto osservare» e «accumulano una gigantesca quantità di conoscenza soltanto da ciò». Queste affermazioni non sono digressioni marginali: esse costituiscono la giustificazione metodologica fondamentale per l'intera strategia dell'apprendimento passivo da video, che è il cuore dell'approccio JEPA al world model. Il documento programmatico «A Path Towards Autonomous Machine Intelligence» (LeCun, 2022) include un grafico dello sviluppo infantile tratto dal lavoro di Emmanuel Dupoux con la didascalia che la maggior parte della conoscenza infantile viene acquisita «prevalentemente tramite osservazione, con pochissima interazione diretta», e questo grafico funge da giustificazione per l'apprendimento autosupervisionato da video passivi.

La contraddizione tra le affermazioni di LeCun sul rapporto tra pensiero e linguaggio e il corpus della filosofia del linguaggio non è una questione di sfumatura o di interpretazione: è una contraddizione diretta tra una posizione empirica e una tradizione di argomenti filosofici ed evidenze empiriche che si estende per secoli. Meriti un'analisi più sistematica che identifichi esattamente dove e perché le premesse di LeCun falliscono, perché questo fallimento ha conseguenze dirette sulla plausibilità dell'architettura che su di esse si fonda.

La tesi di LeCun — che il pensiero non abbia "assolutamente nulla a che fare con il linguaggio" — è nella migliore delle ipotesi una versione estrema e infondata del separatismo cognitivo: la posizione che i processi cognitivi sono fondamentalmente indipendenti dal linguaggio e che il linguaggio serve soltanto come mezzo di espressione per pensieri preformati. Questa posizione ha una storia rispettabile — è associata in forma moderata a Fodor e alla sua ipotesi del "Language of Thought" — ma nemmeno Fodor sosteneva che il pensiero fosse completamente indipendente dal linguaggio naturale appreso. Fodor sosteneva che il pensiero avvenisse in un mentalese interno, ma questo mentalese aveva una struttura proposizionale compositiva che doveva essere acquisita attraverso l'interazione con il linguaggio naturale e con l'ambiente.

Le evidenze empiriche contro il separatismo cognitivo estremo sono schiaccianti. Gli studi di Boroditsky e colleghi hanno dimostrato che la lingua parlata influenza la percezione del colore (le lingue con più distinzioni cromatiche producono parlanti che discriminano i colori più rapidamente), la percezione del tempo (le

lingue con diverse metafore spaziali per il tempo producono concettualizzazioni diverse della temporalità), e il ragionamento causale (le lingue che codificano l'agentività in modo diverso producono diverse attribuzioni di responsabilità). Non si tratta di influenze marginali: sono differenze sistematiche nel ragionamento che si manifestano in compiti non linguistici. Se il pensiero non avesse "nulla a che fare con il linguaggio", queste influenze non esisterebbero.

Il problema è che queste affermazioni contraddicono il corpus quasi intero della filosofia del linguaggio

Il problema è che queste affermazioni contraddicono il corpus quasi intero della filosofia del linguaggio dall'illuminismo tedesco in poi. Johann Georg Hamann, nella sua Metacritica della critica della ragion pura (1784), aveva già indicato che la ragione non può operare indipendentemente dal linguaggio: non esiste cognizione «pura» pre-linguistica, perché il linguaggio non è uno strumento esterno che la ragione adopera per esprimersi, ma la condizione di possibilità stessa del pensiero. Wittgenstein nelle *Ricerche filosofiche* (1953) ha radicalizzato questa intuizione: il significato è l'uso, e i giochi linguistici non codificano un pensiero preesistente ma lo costituiscono. Il significato di «dolore» non è un'entità mentale privata che le parole traducono in simboli pubblici; è determinato dalle regole del gioco linguistico in cui il termine viene impiegato. Vygotsky aveva mostrato empiricamente, attraverso decenni di osservazione dello sviluppo infantile, che le funzioni cognitive superiori si sviluppano attraverso l'interiorizzazione del discorso sociale: il bambino prima parla ad alta voce mentre risolve problemi, poi abbassa la voce a mormorio, poi interiorizza il discorso come monologo interiore silenzioso. La tesi che il pensiero non abbia «nulla a che fare con il linguaggio» è incompatibile con questa tradizione empirica e filosofica di vastissimo respiro.

Sul piano dello sviluppo infantile, la premessa dell'osservazione passiva come meccanismo cognitivo primario è egualmente infondata. I bambini non sono fotocamere: sono agenti incarnati con sistemi sensoriali integrati, con una struttura biologica di attenzione che privilegia i volti umani, il movimento biologico e le discontinuità nell'esperienza, con un sistema motorio in continuo sviluppo che fornisce feedback propriocettivo e aptico che trasformano radicalmente la natura dell'apprendimento percettivo. La «permanenza dell'oggetto», che il Capitolo I di questa monografia richiama correttamente come conquista piagetiana, non emerge dall'osservazione passiva ma dall'interazione motoria: il bambino impara che un oggetto nascosto dietro uno schermo persiste proprio perché può agire su quello schermo, spingerlo, spostarlo, recuperare l'oggetto. L'apprendimento è dialogico e incarnato, non monologico e visivo. Nessuna camera che registra video passivamente può replicare questa struttura.

Perché questa critica delle premesse di LeCun è rilevante per l'architettura tecnica di JEPA? Perché le premesse filosofiche determinano le scelte architetturali, e le scelte architetturali determinano i limiti strutturali dei sistemi. Se la comprensione del mondo fisico è costitutivamente dipendente dall'incarnazione, dall'interazione e dal linguaggio — come la convergenza di prove filosofiche ed empiriche

suggerisce — allora un sistema che apprende esclusivamente da video passivi e non possiede incarnazione né linguaggio non potrà, per ragioni di principio, raggiungere quella comprensione. Il documento V-JEPA 2 (Assran et al., 2025) stesso rivela involontariamente questa tensione quando specifica che, poiché i dati non curati «possono ostacolare le prestazioni del modello», il corpus di addestramento viene filtrato attraverso una pipeline di curazione che utilizza come distribuzione bersaglio i dataset Kinetics, Something-Something v2, COIN e EpicKitchens — tutti dataset costruiti da esseri umani, che codificano giudizi umani su quali azioni sono rilevanti, quali categorie sono discriminative, quale struttura temporale è significativa. Il sistema non apprende dal mondo; apprende da una rappresentazione del mondo già pre-strutturata attraverso categorie umane. Le premesse di LeCun promettono il primo; l'architettura JEPA può al meglio offrire il secondo.

XI.3 — I tre teoremi di impossibilità: dimostrazione formale

La critica alle premesse filosofiche di LeCun, per quanto motivata, non è sufficiente da sola a stabilire un limite strutturale. Per fare ciò è necessario appoggiarsi a risultati matematici dimostrati che vincolano categorialmente ciò che qualsiasi sistema formale può conseguire. Tre teoremi, convergenti nella loro implicazione, definiscono questo perimetro.

XI.3.1 — L'impossibilità di radicamento del riferimento: l'argomento modello-teorico di Putnam

Il primo teorema concerne la relazione di riferimento — il legame tra una rappresentazione e ciò che essa rappresenta. La domanda è se operazioni su rappresentazioni possano *stabilire* tale relazione.

Frege (1892) aveva già distinto *Sinn* (senso, modo di presentazione) da *Bedeutung* (riferimento, l'oggetto denotato): «la stella del mattino» e «la stella della sera» hanno sensi differenti ma lo stesso riferimento (Venere). Un sistema che operi su sensi — che manipoli modi di presentazione — non opera ipso facto sui referenti. La trasformazione di una descrizione in un'altra trasforma descrizioni, non ciò che le descrizioni descrivono.

Putnam (1981) ha formalizzato questa intuizione attraverso la logica matematica. Sia T una teoria del primo ordine coerente e sia M la sua interpretazione intenzionale — il modello che T intende descrivere. Il teorema di Löwenheim-Skolem afferma che se T ha un modello infinito, allora T ha modelli di ogni cardinalità infinita. Il teorema di completezza di Gödel (1930) garantisce che ogni teoria del primo ordine coerente ammette un modello. Putnam ha quindi dimostrato il seguente risultato: per qualsiasi teoria coerente T e qualsiasi interpretazione intenzionale M, esistono interpretazioni non-standard M* tali che (i) M* soddisfa esattamente le stesse frasi di T che M soddisfa, (ii) M* assegna oggetti diversi ai termini di T rispetto a M, e (iii) nessuna proprietà formale di T distingue M da M*.

La dimostrazione procede per permutazione: sia π una permutazione non triviale del dominio di M, e si definisca l'interpretazione di ogni predicato P in M* come {π(x₁), …, π(x) : (x₁, …, x) ∈ P^M}. La permutazione preserva la soddisfazione, dunque M* soddisfa le stesse frasi di M, ma assegna l'oggetto π(c^M) alla costante c, che differisce da c^M ogniqualvolta π è non-triviale. La struttura formale di T — le frasi che dimostra, le inferenze che autorizza — è identica nelle due interpretazioni.

Il corollario è fondamentale: il riferimento non è determinato dalla struttura formale. Qualsiasi sistema formale ammette molteplici interpretazioni che soddisfano i suoi vincoli formali; il formalismo non può determinare quale interpretazione sia «corretta» perché tutte la soddisfano egualmente. Applicato a una rete neurale: il pattern di attivazione che correla con «gatto» nel corpus di addestramento correla egualmente con «immagini-di-gatti-come-array-di-pixel», con «oggetti-etichettati-come-gatti-da-umani», con «qualsiasi proprietà che si allinei con 'gatto' nella distribuzione di addestramento». La rete non può determinare a quale di queste proprietà si «riferisce» perché la sua struttura funzionale è compatibile con tutte le interpretazioni che preservano le mappature input-output sul corpus di addestramento. Il riferimento è sottoderminato dal calcolo.

XI.3.2 — L'impossibilità di acquisizione della normatività: l'argomento di Kripke sulle regole

Il secondo teorema riguarda la normatività — il dominio di come le cose *dovrebbero* essere, non di come semplicemente sono. Comprendere l'addizione significa capire che 2 + 2 *deve* fare 4, che 2 + 2 = 5 è *sbagliato* indipendentemente da quante volte qualcuno lo scriva. L'apprendimento statistico estrae pattern da dati osservati; la domanda è se tale estrazione possa produrre norme.

Kripke (1982) ha dimostrato che nessun insieme finito di esempi determina univocamente una regola. Sia f una qualsiasi funzione calcolabile (ad esempio la funzione addizione +) e sia E = {(x₁, f(x₁)), …, (x , f(x))} un insieme finito di esempi di input-output di f. Il teorema afferma che per qualsiasi insieme finito E e qualsiasi funzione f coerente con E, esistono infinite funzioni calcolabili f* tali che (i) f* coincide con f su tutti gli input che compaiono in E, e (ii) f* diverge da f su qualche input non comparso in E.

La dimostrazione costruisce esplicitamente un controesempio. Si definisca la funzione *quus* ⊕ come segue: x ⊕ y = x + y se x < k e y < k (dove k supera tutti i valori in E), e x ⊕ y = 5 altrimenti. La funzione quus concorda con plus su tutti gli esempi in E (tutti coinvolgono valori inferiori a k) e diverge da plus per input in cui x ≥ k o y ≥ k. L'insieme E è soddisfatto sia da + che da ⊕. Generalizzando: per qualsiasi f e qualsiasi E finito, sia k superiore a tutti i valori in E e si definisca f*(x) = f(x) se x < k, e f*(x) = c (costante arbitraria) altrimenti. f* concorda con f su E, diverge al di fuori di E, e infinite tali f* esistono al variare di c o della soglia k.

Il corollario — che Kripke denomina «divario della normatività» — è questo: l'apprendimento statistico da esempi finiti non può acquisire regole. Un modello esposto a esempi E apprende una funzione coerente con E; infinite funzioni sono coerenti con E; quale di esse il modello generalizzerà a input inediti è determinato dall'architettura e dall'algoritmo di ottimizzazione, non dagli esempi stessi; se quella generalizzazione è «corretta» — se corrisponde alla regola intenzionale — non è determinato da nulla nel processo di apprendimento.

Il punto metafisico più profondo è che non esiste alcun fatto negli esempi riguardo a quale regola essi istanzino. Gli esempi 2, 4, 6, 8 non *contengono* la regola «+2» o la regola «quus-plus-2»; sono compatibili con entrambe. Seguire una regola richiede di cogliere la regola stessa — di comprendere ciò che rende corrette o scorrette le sue applicazioni. Gli esempi da soli non possono fornire questo. Ciò spiega strutturalmente la fragilità delle reti profonde sullo spostamento di distribuzione: il modello non ha imparato regole; ha appreso pattern. I pattern estrapolano secondo l'architettura; le regole si applicano secondo il significato. Il modello non può distinguere i due perché non ha accesso al significato.

XI.3.3 — L'impossibilità di accesso allo spazio delle ragioni: Tarski e Gödel

Il terzo risultato concerne lo «spazio delle ragioni» — il dominio in cui le affermazioni intrattengono relazioni inferenziali, in cui le asserzioni possono essere giustificate, sfidate e rivedute attraverso lo scambio di ragioni (Sellars, 1956; Brandom, 1994). Comprendere un'affermazione significa occupare questo spazio: cogliere da cosa essa segue, cosa da essa segue, e cosa costituirebbe evidenza per o contro di essa.

Il teorema di indefinibilità di Tarski (1936) dimostra che la verità — la proprietà semantica per eccellenza — non può essere definita all'interno della sintassi per linguaggi sufficientemente espressivi. Sia L un linguaggio formale capace di esprimere l'aritmetica elementare e sia T(x) un predicato di L inteso a esprimere «x è vero in L». Il teorema afferma che non esiste in L alcuna formula T(x) tale che per ogni frase φ di L valga: $T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$, dove $\ulcorner \varphi \urcorner$ è il nome o codice di φ in L.

La dimostrazione procede per diagonalizzazione. Si supponga per assurdo che esista tale T(x). Per il lemma diagonale — conseguenza della capacità espressiva di L — esiste una frase λ tale che: $\lambda \leftrightarrow \neg T(\ulcorner \lambda \urcorner)$. Se λ è vera, allora $T(\ulcorner \lambda \urcorner)$ vale (per lo schema T), ma λ asserisce $\neg T(\ulcorner \lambda \urcorner)$, dunque λ è falsa: contraddizione. Se λ è falsa, allora $\neg T(\ulcorner \lambda \urcorner)$ vale, ma $\lambda \leftrightarrow \neg T(\ulcorner \lambda \urcorner)$ implica che λ è vera: contraddizione. Entrambe le assunzioni conducono a contraddizione, dunque non esiste T(x) con le proprietà richieste. La semantica eccede la sintassi.

Il primo teorema di incompletezza di Gödel (1931) rafforza questa conclusione in un modo diverso: per qualsiasi sistema formale coerente F capace di esprimere l'aritmetica, esiste una frase G tale che (i) G è vera (nel modello standard dell'aritmetica), (ii) G non è dimostrabile in F, e (iii) ¬G non è dimostrabile in F. La derivabilità sintattica (dimostrabilità in F) non coincide con la verità semantica

(verità nel modello standard): esistono verità semantiche che nessuna operazione sintattica all'interno di F può stabilire. Il divario è strutturale, non contingente: non si colma con più derivazioni, più calcolo, più manipolazione formale.

Combinando i due risultati con la tradizione sellar-siana-brandominiana dello spazio delle ragioni si ottiene la terza impossibilità: le operazioni sintattiche non possono accedere allo spazio delle ragioni, perché le relazioni costitutive di quello spazio — implicazione, incompatibilità, giustificazione — sono relazioni semantiche tra significati, non relazioni formali tra simboli non interpretati. Un sistema che opera puramente su sintassi non può partecipare alle relazioni inferenziali che costituiscono la comprensione. Interrogato «perché hai classificato questo come cane?», il sistema non può rispondere nello spazio delle ragioni: può fornire una storia causale (questi pesi, queste attivazioni, questo output), ma non una giustificazione (a causa di queste caratteristiche, che costituiscono evidenza per essere-un-cane, dati il concetto di cane e le sue relazioni inferenziali). Il sistema possiede cause; è privo di ragioni.

XI.3.4 — La convergenza delle impossibilità

Le tre impossibilità non sono osservazioni indipendenti. Sono manifestazioni di un unico fatto strutturale: la sintassi non determina la semantica. Ciascuna impossibilità è fondata su teoremi matematici dimostrati — l'argomento di Putnam usa il teorema di Löwenheim-Skolem e il teorema di completezza della logica del primo ordine; l'argomento di Kripke usa la sottoderminazione delle funzioni da esempi finiti, un risultato elementare della teoria della calcolabilità e della filosofia della matematica; il teorema di Tarski è dimostrato per diagonalizzazione e quello di Gödel per aritmetizzazione e auto-riferimento. Questi non sono pareri ma teoremi.

Qualsiasi sistema che operi formalmente — che manipoli rappresentazioni secondo regole, apprenda da esempi finiti attraverso ottimizzazione numerica, calcoli funzioni su strutture discrete — eredita queste limitazioni. V-JEPA 2 è esattamente tale sistema: manipola rappresentazioni secondo funzioni apprese, apprende da esempi finiti attraverso gradient descent, calcola funzioni su array numerici. I risultati di impossibilità si applicano pertanto con piena forza.

XI.4 — Il meccanismo del lavaggio epistemico in V-JEPA 2

Con questo apparato formale è ora possibile descrivere con precisione il meccanismo attraverso cui V-JEPA 2 — come qualsiasi sistema che istanzi il paradigma meccanicistico — oscura sistematicamente la distanza tra ciò che consegue e ciò che afferma di conseguire. Nourizadeh (2026) denomina questo meccanismo «lavaggio epistemico» (epistemic laundering): la struttura entra nel sistema in un punto, scompare dalla vista durante l'elaborazione, e riemerge all'output attribuita erroneamente al sistema come scoperta emergente piuttosto che come preservazione di input ereditato.

Il meccanismo si dispiega in otto stadi riconoscibili e formalmente identificabili.

Nel *primo stadio*, struttura ereditata entra nel sistema. V-JEPA 2 viene addestrato su video che non sono sequenze casuali di pixel: sono artefatti umani, ripresi da esseri umani per scopi umani, montati da editor umani, filtrati da ricercatori umani attraverso una pipeline di curazione esplicita. Come indicato nella sezione 2.3 di Assran et al. (2025), il corpus YT1B viene filtrato attraverso una distribuzione bersaglio composta dai dataset di addestramento di Kinetics, Something-Something v2, COIN e EpicKitchens — dataset che codificano giudizi umani su quali azioni sono rilevanti, quali categorie sono discriminative, quale granularità temporale è significativa. La struttura categoriale non è generata dal sistema; è ereditata.

Nel *secondo stadio*, l'elaborazione formale avviene attraverso macchinari opachi. Il Vision Transformer codifica fotogrammi in rappresentazioni latenti, il predittore genera rappresentazioni di regioni mascherate, il target encoder aggiornato con EMA fornisce i target. Le operazioni sono definite matematicamente con precisione — la perdita si scrive $\mathcal{L}_IJEPA(\theta, \varphi) = \mathbb{E}_{X,M}[\| s_\varphi(E\theta(X_C), M) - sg(\bar{E}_\theta(X_T)) \|^2]$ — e sono semanticamente opache: miliardi di parametri si applicano in sequenza senza che alcun elemento della catena computazionale abbia accesso al significato delle strutture che manipola.

Nel *terzo stadio*, la provenienza viene oscurata. I video vengono descritti come «dati» tout court, non come «video ripresi da umani, montati da umani, selezionati da umani, etichettati da umani». Il vocabolario tecnico — «spazio latente», «embedding», «perdita predittiva» — è accurato per descrivere le operazioni formali ed è sistematicamente silenzioso sulla struttura ereditata in ingresso. I benchmark vengono presentati come misure di prestazione senza notare che i benchmark stessi presuppongono la struttura categoriale umana che si pretende di avere appreso.

Nel *quarto stadio*, gli output vengono erroneamente attribuiti al sistema. Il documento V-JEPA 2 afferma che il sistema «apprende a comprendere e predire come si evolve il mondo fisico» (Assran et al., 2025). L'affermazione attribuisce la comprensione al modello. Un'affermazione più accurata sarebbe: «V-JEPA 2 apprende ad approssimare una funzione da input strutturati da umani a output definiti da umani che era implicitamente contenuta nei dati strutturati da umani su cui è stato addestrato».

Nel *quinto stadio*, i practitioner non riescono ad articolare cosa il sistema comprenda. Come si è analizzato nella sezione XI.2, LeCun sostiene che i suoi sistemi non comprendono perché mancano ancora di elementi architetturali aggiuntivi — world model integrato, pianificazione gerarchica — senza riconoscere che la premessa filosofica su cui l'intera architettura si fonda è empiricamente infondata e teoricamente confutata. Non riesce ad articolare cosa il sistema comprenda perché non c'è nulla da articolare: la struttura categoriale che emerge all'output è la medesima struttura categoriale umana che è entrata all'input.

Nel *sesto stadio*, vengono dispiegati pseudo-spiegazioni. Quando interrogato sull'assenza di comprensione genuina, l'ecosistema intellettuale del paradigma JEPA-AMI invoca «l'emergenza alla scala». Il claim è strutturalmente tautologico: afferma che se si fa abbastanza di ciò che non ha funzionato, funzionerà. Non fornisce alcun meccanismo per la trasformazione qualitativa sperata. Come si è dimostrato nella sezione XI.3, i vincoli di impossibilità sono categoriali, non quantitativi: un motore sintattico più grande rimane un motore sintattico; più corrispondenza di pattern resta corrispondenza di pattern. L'invocazione dell'emergenza maschera questa irriducibilità categoriale suggerendo che un aumento quantitativo possa produrre un cambiamento categoriale — suggestione priva di qualsiasi fondamento.

Nel *settimo stadio*, il frame problem si manifesta. V-JEPA 2 non può determinare la rilevanza dall'interno delle proprie operazioni. Il sistema può processare soltanto ciò che il suo corpus di addestramento lo ha preparato a processare; input fuori distribuzione producono comportamenti arbitrari; il sistema non ha accesso a criteri di rilevanza che gli permettano di identificare quando gli input eccedono la propria competenza. La curazione dei dati — necessità ammessa esplicitamente nel documento tecnico — è la prova puntuale di questo limite: senza la pre-selezione umana di ciò che è rilevante, il sistema non funziona.

Nell'*ottavo stadio*, la cecità di secondo ordine diventa operativa. I fallimenti del sistema vengono sistematicamente ridenominati come virtù attraverso un vocabolario tecnico che neutralizza la valenza negativa. La «black box problem» diventa «complessità fondamentale». L'incapacità di spiegare le proprie decisioni diventa «capacità emergente». Il catastrophic forgetting diventa «sfida dell'apprendimento continuo». Le allucinazioni diventano «generazione creativa». L'assenza di comprensione diventa «comprensione emergente in prospettiva». In ciascun caso, il fallimento viene ridenominato con un termine che porta valenza positiva, rendendo invisibile la pressione verso una soluzione genuina.

XI.5 — La cecità di secondo ordine e il meccanismo di conversione delle anomalie

Il concetto di «cecità di secondo ordine» merita una trattazione separata perché identifica un fenomeno epistemologico di grande portata che va ben oltre il caso specifico di V-JEPA 2. La cecità di primo ordine sarebbe il fallimento di vedere qualcosa. La cecità di secondo ordine è il fallimento di vedere che si sta fallendo di vedere — l'invisibilità dell'invisibilità. Essa opera attraverso un meccanismo di conversione delle anomalie in virtù che segue un pattern ricorrente e formalmente identificabile.

Kuhn (1962) aveva identificato le anomalie come fonte di crisi scientifica: un'anomalia è un fenomeno che resiste alla spiegazione all'interno del paradigma, un puzzle che non può essere risolto con le risorse concettuali disponibili. L'analisi kuhniana presuppone però che le anomalie vengano riconosciute come tali — che i fallimenti vengano registrati come fallimenti. La cecità di secondo ordine emerge

precisamente quando un paradigma sviluppa meccanismi per impedire tale riconoscimento. Non si tratta di negazione deliberata — i practitioner spesso *credono genuinamente* che le anomalie siano virtù — ma di un processo che opera al livello del vocabolario e delle categorie di descrizione, al di sotto della soglia della scelta cosciente.

Il meccanismo di conversione segue quattro passi. Anzitutto, un fallimento si manifesta: il sistema non fa ciò che afferma o ci si aspetta faccia. Poi, il fallimento viene nominato: un termine viene introdotto che descrive il fenomeno in modo neutro. Successivamente, il nome viene sostituito: un nuovo termine viene introdotto che descrive lo stesso fenomeno con valenza positiva. Infine, la sostituzione diventa standard: il nuovo termine viene adoperato sistematicamente, il vecchio termine decade, e la discussione procede come se il fenomeno descritto dal nuovo termine fosse qualitativamente distinto da quello descritto dal vecchio.

Nel caso di V-JEPA 2 e dell'apprendimento profondo in generale, questa conversione è sistematica e quasi esaustiva. L'opacità delle rappresentazioni interne diventa «complessità ad alta dimensionalità». L'inesplicabilità degli output diventa «capacità emergenti». L'incapacità di accumulare conoscenza senza catastrophic forgetting diventa «sfida dell'apprendimento continuo». La produzione di affermazioni false con la stessa sicurezza di quelle vere diventa «generazione creativa» o, con una denominazione che neutralizza addirittura il riferimento alla falsità, «allucinazione». L'assenza documentata di comprensione nel senso strutturalmente forte viene deferita nel tempo come «comprensione che emergerà alla scala».

La funzione epistemica di queste conversioni non è ornamentale: è la sopravvivenza del paradigma in presenza di evidenza contraria. La conversione non aggiunge informazione — tutti sanno che i sistemi di deep learning soffrono di catastrophic forgetting, che allucinano, che mancano di comprensione. Ciò che essa modifica è il modo in cui queste informazioni vengono *vissute* — se esse vengono sperimentate come problemi che creano pressione verso soluzioni o come caratteristiche di sistemi profondi che nessun programma di ricerca superficiale potrebbe avere. La cecità di secondo ordine non opera sulla realtà dei fallimenti ma sulla loro valenza epistemica.

Lakatos (1978) aveva identificato il segno distintivo di un programma di ricerca degenerante: le modifiche ad hoc che accomodano le anomalie senza generare nuove predizioni. Le conversioni sopra descritte sono, nel senso più preciso, modificazioni ad hoc che operano a livello lessicale piuttosto che teorico. Il teorema non cambia; la descrizione del teorema cambia. I fallimenti permangono; non vengono più denominati come tali. La differenza è che operando al livello del vocabolario piuttosto che della teoria, queste conversioni eludono i criteri lakatosiani di valutazione: come si potrebbe valutare un programma come degenerante quando i suoi fallimenti sono stati lessicalmente trasformati in virtù?

XI.5bis — GMJE e i limiti del lavaggio epistemico: una valutazione comparativa

L'analisi del meccanismo del lavaggio epistemico sviluppata nella sezione XI.4 può essere applicata al framework GMJE di Huang (2026) per valutarne il posizionamento rispetto ai sistemi precedentemente analizzati. La questione non è se GMJE superi i limiti strutturali identificati dai teoremi di Putnam, Kripke e Tarski — essendo un sistema formale, non può — ma se il suo contributo tecnico modifichi significativamente il confine tra conseguimento genuino e affermazione eccessiva.

Sul lato del conseguimento genuino, GMJE produce tre avanzamenti rispetto a I-JEPA che non sono riducibili al semplice incremento quantitativo. Primo: la derivazione formale del Mahalanobis Trace Trap (Appendice O) dimostra algebricamente perché la NLL empirica collassa in una costante, trasformando una difficoltà pratica nota in un risultato negativo dimostrato. Secondo: il teorema InfoNCE bridge (sezione 4.6) prova che l'apprendimento contrastivo è il limite degenere di GMJE con $\Sigma_ct = 0$, unificando due famiglie architetturali prima separate in un quadro teorico comune. Terzo: la disponibilità di una densità generativa esplicita p(z_t) in forma chiusa è una proprietà strutturale nuova che gli esperimenti di latent sampling su MNIST (sezione 5.3) mostrano avere valore empirico misurabile: la distanza media tra campioni latenti GMJE e cluster semantici reali è inferiore a quella dei metodi baselines (Fig. 9c di Huang, 2026).

Sul lato delle affermazioni potenzialmente eccessive, il lavaggio epistemico può manifestarsi se si afferma che GMJE 'modella la struttura multimodale del mondo' in un senso che trascende il pattern-matching distribuito. Il Teorema di Löwenheim-Skolem si applica a GMJE esattamente come a I-JEPA: i prototipi μ_k della mistura sono coerenti con infinite interpretazioni della struttura categoriale del mondo, e nessun criterio formale interno al sistema discrimina tra esse. L'argomento di Kripke sulla normatività si applica ai pesi di routing $\gamma_k(z_c)$: nessun insieme finito di esempi determina univocamente quali regole di routing il sistema 'dovrebbe' seguire per input out-of-distribution. La Stanza Cinese di Searle si applica all'intera pipeline GMJE-MDN: il bottleneck informativo (la rete parametrica osserva solo z_c) è un vincolo architetturale motivato dal rischio di identity collapse, non dall'acquisizione di comprensione semantica genuina.

L'isomorfismo con la teoria delle stringhe identificato da Nourizadeh (2026) si ripete, con le debite proporzioni, in GMJE: l'eleganza matematica del framework — la chiusura delle marginali GMM, la derivazione analitica delle condizionali, l'unificazione di contrastive e predictive SSL sotto un unico principio — non implica che il sistema acceda alla semantica che le strutture matematiche intendono rappresentare. La sofisticazione della struttura formale e la profondità della comprensione semantica rimangono grandezze ortogonali, separabili con precisione attraverso i teoremi di impossibilità della sezione XI.3.

La valutazione complessiva coloca GMJE nel gruppo dei contributi che espandono il confine del conseguimento genuino senza modificare il confine

dell'impossibilità strutturale. Il framework risolve problemi tecnici reali e importanti — la predizione in regimi multimodali, la calibrazione dell'incertezza, il campionamento generativo — senza affrontare i problemi di radicamento, normatività e accesso allo spazio delle ragioni che i teoremi identificano come irrisolvibili dall'interno del paradigma computazionale. Questa conclusione è coerente con il giudizio differenziato della sezione X.4: GMJE sposta il confine tra 'sì nel senso funzionale debole' e 'sì nel senso funzionale forte' verso il secondo, avvicinandosi a rappresentazioni che supportano ragionamento calibrato sull'incertezza, senza però attraversare il confine verso la comprensione nel senso intenzionale o fenomenologico.

XI.6 — Il nucleo inmetabolizzabile: il calcolo come assunzione costitutiva

Il motivo per cui il paradigma computazionale non può trasformarsi internamente per affrontare i problemi che esso stesso genera risiede in una struttura profonda che Nourizadeh (2026) denomina il «nucleo inmetabolizzabile»: l'assunzione costitutiva che definisce il campo e che, per questa ragione, non può essere messa in questione dall'interno senza dissolvere il campo stesso.

Per V-JEPA 2 e per il framework AMI più in generale, questo nucleo è la tesi computazionale sulla cognizione: l'assunzione che la cognizione sia calcolo, che il calcolo sia manipolazione formale di simboli (o, nella variante connessionista, trasformazione numerica secondo funzioni apprese), e che sufficienti risorse computazionali convergano verso l'intelligenza in senso pieno. L'architettura di V-JEPA 2 — come ogni rete neurale profonda — *è* il suo calcolo. L'encoder è una funzione; il predittore è una funzione; il target encoder è una funzione. Rimuovere il calcolo da V-JEPA 2 non significa modificare V-JEPA 2: significa annullarlo.

Questa costitutività ha implicazioni profonde. Il framework AMI non può seriamente interrogarsi se la cognizione sia computazionale, perché tale interrogazione dissolverebbe il framework stesso. Non può prendere sul serio le critiche di Dreyfus (1972, 1992) alla tesi computazionale, perché tali critiche non offrono un «approccio alternativo all'IA» bensì mettono in questione se il soggetto di indagine dell'IA esista nella forma in cui il campo lo presuppone. Non può integrare genuinamente le intuizioni dell'embodied cognition (Varela, Thompson, Rosch, 1991), che non è una variante computazionale dell'IA ma una sfida alla sua assunzione fondamentale. Le alternative — l'approccio enattivo, la psicologia ecologica gibsoniana, la computazione morfologica — vengono sistematicamente marginalizzate non perché siano state confutate ma perché pongono domande che il paradigma non può porre senza dissolversi.

La logica del nucleo inmetabolizzabile si ritrova in modo esatto nel campo adiacente della teoria delle stringhe, che costituisce il caso parallelo diagnosticamente rilevante. La teoria delle stringhe non può interrogarsi sui fondamenti della meccanica quantistica perché essa è la meccanica quantistica applicata a oggetti estesi unidimensionali: ogni risultato celebrato — la

cancellazione delle anomalie di Green-Schwarz (1984), la rete di dualità di Witten (1995), la corrispondenza AdS/CFT di Maldacena (1998) — presuppone la meccanica quantistica standard a ogni livello. Modificare la meccanica quantistica non produrrebbe una «teoria delle stringhe modificata»; produrrebbe una teoria diversa, costruita da zero, che non erediterebbe alcuno dei risultati che giustificano i cinquant'anni di investimento nel programma. Come Witten stesso ammette — affermando che il campo è ancora «largely in the dark» riguardo ai suoi princìpi unificatori e che il suo risultato centrale è «scritto in un linguaggio che non comprendiamo» — la struttura formale del programma ha ragiunto una sofisticazione straordinaria senza che essa si sia tradotta in comprensione fisica nel senso semanticamente pieno del termine.

La convergenza di queste due situazioni — in fisica fondamentale e in intelligenza artificiale, in campi con storie, metodologie e practitioner diversi — non è coincidenziale. Essa rivela un limite del paradigma meccanicistico che è comune a entrambi i domini: la presupposizione che operazioni formali su strutture discrete possano fondare contenuto semantico. Quando questa presupposizione incontra fenomeni in cui la prospettiva è costitutiva — fenomeni che coinvolgono riferimento, normatività, comprensione — il meccanismo del lavaggio epistemico entra in azione e la cecità di secondo ordine ne preserva l'invisibilità.

XI.7 — Ciò che rimane: il confine tra risultati genuini e affermazioni eccessive

La diagnosi fin qui sviluppata non è nichilistica. Essa non afferma che V-JEPA 2 sia inutile né che il framework AMI sia privo di valore. Afferma qualcosa di più preciso e tecnicamente consequenziale: che esiste un confine nitido tra ciò che questi sistemi conseguono genuinamente e ciò che affermano di conseguire, e che riconoscere con precisione questo confine è una condizione necessaria tanto per l'onestà intellettuale quanto per la possibilità di progresso genuino.

Sul lato del conseguimento genuino si collocano capacità di riconoscimento di pattern in domini vincolati di straordinaria qualità. V-JEPA 2 apprende rappresentazioni video semanticamente ricche che supportano classificazione di azioni, previsione di eventi e trasferimento di rappresentazioni verso task supervisionati. Questi sono risultati reali, misurabilmente superiori agli approcci alternativi sulle distribuzioni di valutazione considerate. Il sistema funziona nel senso ingegneristico del termine: produce output utili per applicazioni definite. I risultati di LeJEPA relativi alla distribuzione gaussiana isotropica come ottimo per il probing lineare (Teorema 1 della sezione III.5bis di questa monografia) sono risultati matematicamente dimostrati che modificano il paesaggio teorico dell'apprendimento autosupervisionato. Il meccanismo DAAM per il parlato (Capitolo VII) produce token a una frequenza inferiore rispetto ai codec neurali comparabili mantenendo la reversibilità perfetta dell'operazione di codifica.

Sul lato delle affermazioni eccessive si collocano le pretese di «comprensione del mondo fisico», di «modelli del mondo» nel senso di modelli causali capaci di

supportare ragionamento controfattuale, e di architetture che si approssimano all'intelligenza generale. Queste affermazioni oltrepassano ciò che i sistemi consegnano, e lo oltrepassano non in modo contingente — non perché manchino ancora dati, parametri o innovazioni architetturali — ma per le ragioni strutturali dimostrate nelle sezioni XI.3 e XI.4.

Il vocabolario adoperato per descrivere questi sistemi compie lavoro epistemico non neutro. «Apprendere», «comprendere», «mondo model» implicano acquisizione di conoscenza, accesso al significato, rappresentazione di strutture causali del mondo — tutte proprietà che i tre teoremi di impossibilità dimostrano non poter essere fondate attraverso manipolazione formale di strutture discrete. Usare tale vocabolario per descrivere sistemi che approssimano funzioni da input strutturati da umani a output definiti da umani non è imprecisione accidentale: è il meccanismo lessicale attraverso cui il lavaggio epistemico si perpetua.

Una conseguenza pratica diretta riguarda le decisioni di impiego. I sistemi di riconoscimento di pattern funzionano bene dove il riconoscimento di pattern è sufficiente: sistemi di raccomandazione, classificazione di immagini in domini vincolati, riconoscimento vocale in condizioni controllate, navigazione autonoma in ambienti strutturati. Essi non funzionano, e non possono funzionare per le ragioni di principio fin qui articolate, dove è richiesta comprensione genuina: diagnosi medica che esige giudizio clinico, decisioni giuridiche che richiedono valutazione normativa, contesti educativi che presuppongono comprensione pedagogica, qualsiasi dominio in cui la responsabilità — intesa come capacità di dare ragioni e di essere tenuto a rispondere di esse — sia costitutiva della funzione. Impiegare sistemi che mancano strutturalmente di comprensione in questi domini sulla base di affermazioni di comprensione è una forma di negligenza epistemica che le dimostrazioni di questa sezione rendono non più difendibile come mero eccesso di entusiasmo.

XI.8 — Implicazioni per il programma di ricerca AMI

Le analisi sviluppate in questo capitolo non invalidano il programma di ricerca AMI; lo riorientano. Il framework di LeCun identifica correttamente componenti funzionali che un sistema intelligente deve possedere — percezione, modellazione del mondo, pianificazione, memoria — e offre intuizioni architetturali valide su come questi componenti possano essere integrati. Ciò che il framework non può offrire — e che le premesse filosofiche su cui si fonda incorrettamente promettono — è la colmatura del divario tra elaborazione formale e comprensione semantica attraverso il solo aumento di scala o raffinamento architetturale.

Il progresso genuino in direzione dell'intelligenza generale richiede, sul piano della ricerca fondamentale, di interrogare l'assunzione costitutiva che la cognizione sia esauribile attraverso il calcolo. Questo non significa abbandonare il calcolo come strumento — i risultati matematici e ingegneristici documentati in questa monografia sono reali e preziosi — ma riconoscere il calcolo come strumento tra altri, non come fondamento esaustivo della cognizione. Significa prendere sul

serio, con investimento istituzionale e non soltanto di maniera, le tradizioni di ricerca che hanno messo in questione questa assunzione: l'embodied cognition di Varela, Thompson e Rosch; la psicologia ecologica di Gibson; gli approcci enattivi; la critica di Dreyfus; le teorie causali del contenuto di Dretske e Fodor che questa monografia richiama nel Capitolo IX senza tuttavia articolarne le implicazioni architetturali. Queste non sono alternative «vecchio stile» superate dalla scala dei moderni sistemi: sono programmi di ricerca che pongono domande che il paradigma computazionale non può porre senza dissolversi.

Sul piano metodologico, le implicazioni sono altrettanto precise. I benchmark di valutazione dovrebbero essere progettati in modo da distinguere la corrispondenza di pattern — ciò che i sistemi attuali conseguono realmente — dalla comprensione — ciò che i sistemi attuali non conseguono. La differenza non è meramente quantitativa; è strutturale. Un sistema che raggiunge il 95% di accuratezza su un benchmark di riconoscimento di azioni non ha il 95% di comprensione delle azioni umane: ha appreso ad approssimare con alta fedeltà la funzione implicita nel corpus di addestramento. Questi sono risultati diversi che richiedono descrizioni diverse e che implicano capacità di trasferimento diverse.

Infine, le premesse filosofiche che strutturano le scelte architetturali dovrebbero essere sottoposte a verifica critica esplicita prima di essere adottate come fondamento di programmi di ricerca ad ampia scala. La tesi che il pensiero non abbia «nulla a che fare con il linguaggio» non è un'ovvietà: è una posizione filosofica contestata e, alla luce della tradizione da Hamann a Wittgenstein a Vygotsky, mal fondata. Adottarla implicitamente come premessa di un'architettura significa costruire su fondamenta che il corpus della ricerca filosofica pertinente ha già mostrato come inadeguate. La saggezza metodologica che questo capitolo invoca non è il conservatorismo scettico né l'entusiasmo acritico, ma la combinazione di rigore tecnico, onestà epistemica sui limiti dimostrabili, e sensibilità filosofica ai problemi che la tecnica da sola non può vedere — la medesima triplice disposizione che, nella sezione X.6 di questa monografia, si è dichiarata come l'ideale metodologico dell'intera opera.

Riferimenti integrativi al presente capitolo: Assran et al. (2025); Brandom (1994); Dreyfus (1972, 1992); Frege (1892); Gödel (1930, 1931); Green & Schwarz (1984); Greene & Witten (2025); Hamann (1784); Kripke (1982); Kuhn (1962); Lakatos (1978); LeCun (2022, 2024); Maldacena (1998); Nourizadeh (2026); Putnam (1981); Sellars (1956); Tarski (1936); Varela, Thompson & Rosch (1991); Vygotsky (1934/1986); Witten (1995); Wittgenstein (1953).

XI.9 — Verso un'epistemologia dell'intelligenza artificiale: proposta di una ricerca critica integrata

La critica sviluppata in questo capitolo non mira a paralizzare la ricerca sull'intelligenza artificiale ma a riorientarla. Identificare con precisione i limiti

strutturali di un paradigma è il prerequisito per progettare un programma di ricerca capace di superarli — o, almeno, di sapere con chiarezza dove si trova il confine tra ciò che è raggiungibile all'interno del paradigma e ciò che richiede un salto concettuale.

Proponiamo qui un'agenda di ricerca critica che emerge dall'analisi dei capitoli precedenti, organizzata intorno a tre assi.

Asse 1: Fondamenti empirici delle premesse filosofiche. Le premesse filosofiche di LeCun — che il pensiero non abbia nulla a che fare con il linguaggio, che i neonati apprendano esclusivamente per osservazione passiva — meritano test empirici diretti nel contesto dell'apprendimento automatico, non soltanto critiche filosofiche. Si potrebbe progettare un esperimento controllato che confronta due sistemi identici in architettura e dati, il primo addestrato su video passivi e il secondo addestrato con un segnale interattivo che mimi l'accoppiamento sensorimotorio infantile. Se le premesse di LeCun fossero corrette, i due sistemi dovrebbero produrre rappresentazioni di qualità comparabile; se la tradizione dell'embodied cognition ha ragione, il secondo dovrebbe superare il primo su compiti che richiedono comprensione causale. Questo esperimento non è stato eseguito — una lacuna significativa nella letteratura empirica.

Asse 2: Sviluppo di metriche che distinguano competenza da comprensione. La maggior parte dei benchmark attuali misura la competenza — la performance su distribuzioni simili al training — non la comprensione nel senso strutturale. Sono necessari benchmark che misurino specificamente la generalizzazione composizionale (la capacità di combinare concetti mai visti insieme durante il training), il ragionamento causale controfattuale (la capacità di predire le conseguenze di interventi su variabili causali), e la robustezza ai bias di dataset (la capacità di produrre output corretti quando la distribuzione dei dati è sistematicamente distorta). Queste metriche potrebbero rilevare la distanza tra prestazioni apparenti e comprensione genuina in modo più diretto dei benchmark standard.

Asse 3: Integrazione tra approcci computazionali e approcci embodied. La dicotomia tra AI computazionale e robotica embodied è in parte artificiale e storicamente contingente. Esistono approcci ibridi — come i sistemi di apprendimento per rinforzo con World Model applicati a robot fisici, o i sistemi di imitation learning da dimostrazioni fisiche — che combinano la potenza computazionale dei modelli neurali con il radicamento causale dell'interazione fisica. Investire sistematicamente in questi approcci ibridi, studiandone le rappresentazioni attraverso gli strumenti di JEPA e LeJEPA, potrebbe rivelare se il radicamento causale produce il tipo di miglioramento strutturale nelle rappresentazioni che la tradizione filosofica prevede — o se la correlazione è più debole di quanto la teoria suggerirebbe.

La monografia che avete letto è, in questo senso, un tentativo di fare epistemologia dell'intelligenza artificiale: di applicare alla ricerca sull'IA gli stessi standard di

rigore, onestà e autocritica che quella ricerca applica ai problemi che studia. Non è un tentativo di fermare la ricerca ma di renderla più consapevole di ciò che fa, di ciò che può fare, e di ciò che, per ragioni di principio che i migliori teoremi della matematica e della logica indicano, potrebbe non poter fare.

Appendici

Appendice A — Notazione e Simboli

La presente appendice raccoglie la notazione adoperata sistematicamente nella monografia. L'obiettivo è offrire un riferimento unificato che compensi la stratificazione della notazione attraverso capitoli scritti in momenti diversi.

A.1 — Spazi e insiemi

$\mathcal{X}$ indica lo spazio delle osservazioni. Per immagini, $\mathcal{X} = \mathbb{R}^{H \times W \times C}$ dove H, W sono la risoluzione in altezza e larghezza e C il numero di canali cromatici. Per video, $\mathcal{X} = \mathbb{R}^{T \times H \times W \times 3}$. Per sequenze audio, $\mathcal{X} = \mathbb{R}^{L \times F}$ dove L è la lunghezza temporale e F la dimensione delle feature. Per sequenze di testo, $\mathcal{X}$ è lo spazio delle sequenze di token su un vocabolario finito $\mathcal{V}$ con $|\mathcal{V}|$ elementi.

$\mathcal{Z} = \mathbb{R}^k$ indica lo spazio latente delle rappresentazioni, con $k \ll \dim(\mathcal{X})$.

$\mathcal{A}$ indica lo spazio delle azioni per sistemi agentici. Può essere continuo ($\mathcal{A} \subseteq \mathbb{R}^{d_a}$) o discreto ($\mathcal{A} = \{1, \ldots, |\mathcal{A}|\}$).

$\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ indicano rispettivamente insiemi dei naturali, interi, razionali, reali e complessi.

A.2 — Vettori, matrici e tensori

Vettori colonna: grassetto minuscolo $\mathbf{x} \in \mathbb{R}^d$. Matrici: grassetto maiuscolo $\mathbf{W} \in \mathbb{R}^{m \times n}$. Tensori di ordine ≥ 3: indicizzazione esplicita $x_{i_1, \ldots, i_k}$.

$\|\cdot\|$ senza subscript indica la norma euclidea $\|\mathbf{x}\|_2 = \sqrt{(\Sigma_i x_i^2)}$. $\|\cdot\|_F$ *indica la norma di Frobenius* $\|\boldsymbol{A}\|_F = \sqrt{(\Sigma_{i,j} A_{ij}^2)}$. $\|\cdot\|_{op}$ indica la norma operatore $\|\mathbf{A}\|_{op} = \sup_{\|x\|=1} \|\mathbf{Ax}\|$.

$x \mapsto f(x)$ indica la funzione f applicata all'argomento x. $f \circ g$ indica la composizione $f(g(\cdot))$.

A.3 — Probabilità e statistica

$\mathbb{E}_p[f(X)]$ indica il valore atteso di f(X) sotto la distribuzione p. Quando la distribuzione è omessa o chiara dal contesto si scrive $\mathbb{E}[f(X)]$.

$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$ indica la varianza. $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top]$ è la matrice di covarianza.

$p(x)$ o $p_\theta(x)$ indicano densità di probabilità o funzioni di massa. $p(x \mid y)$ è la densità condizionale.

$\mathcal{N}(\mu, \Sigma)$ indica la distribuzione gaussiana multivariata con media μ e covarianza Σ. $\mathcal{N}(0, I_k)$ è la gaussiana standard isotropica in $\mathbb{R}^k$.

$H(X) = -\mathbb{E}[\log p(X)]$ è l'entropia di Shannon. $H(X \mid Y) = \mathbb{E}_Y[H(X \mid Y = y)]$ è l'entropia condizionale.

$I(X; Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$ è l'informazione mutua.

$KL(p \| q) = \mathbb{E}_p[\log p(X) - \log q(X)]$ è la divergenza Kullback-Leibler da p a q.

A.4 — Architetture neurali

E_θ o φ_θ indicano encoder con parametri θ. P_φ o s_φ indicano predittori con parametri φ. $\bar{E}_{\bar{\theta}}$ indica il target encoder con parametri $\bar{\theta}$ aggiornati via EMA.

$\mathrm{sg}(\cdot)$ indica l'operazione stop-gradient, che propaga il valore in avanti ma blocca il gradiente nella backpropagation: $\partial \mathrm{sg}(\mathbf{x})/\partial \mathbf{x} = \mathbf{0}$, $\mathrm{sg}(\mathbf{x}) = \mathbf{x}$.

$\mathrm{EMA}(\theta, \tau)$ indica l'aggiornamento esponenziale mobile: $\bar{\theta} \leftarrow \tau \cdot \bar{\theta} + (1 - \tau) \cdot \theta$.

MHSA(H) indica la Multi-Head Self-Attention. LN(·) indica Layer Normalization. FFN(·) indica il Feed-Forward Network.

A.5 — Grandezze specifiche delle architetture JEPA

d: dimensione delle rappresentazioni latenti del context encoder (tipicamente 768 per ViT-B, 1024 per ViT-L, 1280 per ViT-H).

d_{pred}: dimensione del predittore, con $d_{pred} < d$ (tipicamente 384).

P: dimensione di una patch in pixel (tipicamente 16).

N: numero totale di patch per un'immagine o un frame ($N = (H/P) \times (W/P)$).

N_C, N_T: numero di patch contestuali e target rispettivamente, con $N_C + N_T \leq N$.

τ: costante di decadimento EMA (tipicamente in $[0.996, 1.0]$).

K: numero di blocchi target per immagine (tipicamente 4).

A.6 — Grandezze specifiche dei World Models

$h_t \in \mathbb{R}^{d_h}$: stato deterministico (recurrent state) al passo t.

$z_t \in \mathbb{R}^{d_z}$: stato stocastico latente al passo t.

$a_t \in \mathcal{A}$: azione al passo t.

$o_t \in \mathcal{X}$: osservazione al passo t.

$p_\theta(z_t \mid h_t)$: prior sullo stato stocastico.

$q_\varphi(z_t \mid h_t, o_t)$: posteriore approssimata sullo stato stocastico.

$f_\theta(h_{t-1}, z_{t-1}, a_{t-1})$: funzione di transizione del recurrent state (GRU o simile).

β: peso del termine KL nella perdita ELBO (KL balancing parameter).

Appendice E — Dimostrazioni GJE/GMJE

E.1 — Dimostrazione del Mahalanobis Trace Trap nel caso unimodale

Sia $Z \in \mathbb{R}^{N \times d}$ una matrice di N vettori di embedding, e sia $C = (1/N)\, Z^T Z \in \mathbb{R}^{d \times d}$ la covarianza empirica calcolata dallo stesso batch. Il termine di data-fit nella NLL è $L_{Maha} = (1/N) \sum_{i=1}^{N} z_i^T C^{-1} z_i$. Per la proprietà $Tr(ab^T) = b^T a$ (uno scalare è uguale alla propria traccia) e per la ciclicità $Tr(AB) = Tr(BA)$:

$$L_{Maha} = (1/N) \sum_i Tr(z_i^T C^{-1} z_i) = (1/N) \sum_i Tr(C^{-1} z_i z_i^T) = Tr(C^{-1} (1/N) \sum_i z_i z_i^T) = Tr(C^{-1} C) = Tr(I_d) = d.$$

Il gradiente $\partial L_{Maha}/\partial\theta = \partial d/\partial\theta = 0$ per qualsiasi parametrizzazione θ dell'encoder. Il termine di data-fit non propaga alcun segnale di apprendimento. L'ottimizzatore rimane con il solo termine $+(1/2) \log|C|$, che viene minimizzato collassando $C \to 0$ (tutte le rappresentazioni verso l'origine). Il collasso dimensionale è il risultato inevitabile. □

E.2 — Dimostrazione dell'InfoNCE bridge

Si parte dalla formulazione DaM-GMJE con M componenti equiprobabili e covarianza isotropica block-diagonale $\Sigma = \mathrm{diag}(\tau I_d, \tau I_d)$. La densità congiunta è: $p(z_c, z_t) = (1/M) \sum_m \mathcal{N}(z_c \mid z_c^{(m)}, \tau I)\, \mathcal{N}(z_t \mid z_t^{(m)}, \tau I)$. Per la regola di Bayes, la NLL condizionale è: $-\log p(z_t \mid z_c) = -\log p(z_c, z_t) + \log p(z_c)$.

Sviluppando $\mathcal{N}(z_c \mid z_c^{(m)}, \tau I)$ con $\|z_c\| = \|z_c^{(m)}\| = 1$ (normalizzazione L2) si ottiene: $\mathcal{N}(z_c \mid z_c^{(m)}, \tau I) = C_0 \exp(z_c^T z_c^{(m)} / \tau)$, con $C_0 = (2\pi\tau)^{-d/2} \exp(-1/\tau)$. Nel limite $\tau \to 0$ (bassa temperatura), il termine del numeratore di Bayes è dominato dal solo indice m^* tale che $z_t = z_t^{(m^*)}$ (coppia positiva), mentre tutti i termini $m \neq m^*$ decadono esponenzialmente. Sostituendo e semplificando le costanti C_0, la NLL condizionale diventa:

$$-\log p(z_t \mid z_c) = -\log \left[\exp(z_c^T z_t / \tau) / \sum_{m=1}^{M} \exp(z_c^T z_t^{(m)} / \tau)\right] + C$$

che è esattamente la perdita InfoNCE di van den Oord et al. (2018), utilizzata in SimCLR, MoCo e tutti i principali metodi contrastivi. Il termine C è una costante

additiva che non influenza l'ottimizzazione. La derivazione dimostra che l'apprendimento contrastivo standard è un caso speciale degenere di GMJE, ottenuto imponendo: (1) covarianza isotropica e block-diagonale ($\Sigma_{ct} = 0$), (2) uguale probabilità a priori per tutti i componenti ($\pi_m = 1/M$), (3) embedding normalizzati sulla sfera unitaria. □

E.3 — Proprietà generative della mistura marginale
Sia la distribuzione congiunta GMJE $p(z_c, z_t) = \Sigma_k \pi_k \mathcal{N}([z_c;z_t] \mid [\mu_{c,k};\mu_{t,k}], \Sigma_k)$. Per il lemma di marginalizzazione delle gaussiane miste, integrando su z_c:

$p(z_t) = \int p(z_c, z_t)\, dz_c = \Sigma_k \pi_k \int \mathcal{N}([z_c;z_t] \mid [\mu_{c,k};\mu_{t,k}], \Sigma_k)\, dz_c = \Sigma_k \pi_k \mathcal{N}(z_t \mid \mu_{t,k}, \Sigma_{tt,k})$,

dove $\Sigma_{tt,k}$ è il blocco diagonale inferiore di Σ_k. La marginalizzazione di una GMM congiunta restituisce una GMM con gli stessi pesi π_k e i parametri marginali $(\mu_{t,k}, \Sigma_{tt,k})$. Il campionamento da $p(z_t)$ avviene in due passi: (1) $k \sim \text{Categorical}(\pi_1,...,\pi_K)$, (2) $z_t \sim \mathcal{N}(\mu_{t,k}, \Sigma_{tt,k})$ via $\Sigma_{tt,k} = A_k A_k^T$ (Cholesky): $z_t = \mu_{t,k} + A_k \varepsilon$, $\varepsilon \sim \mathcal{N}(0,I)$. La complessità computazionale è $O(K d^2)$ per le K decomposizioni di Cholesky, trascurabile rispetto al forward pass dell'encoder. □

Appendice B — Dimostrazioni dei Teoremi Principali

B.1 — Il teorema di ottimalità della distribuzione gaussiana isotropica (LeJEPA, Teorema 1)

Enunciato. Sia $z \in \mathbb{R}^k$ un vettore di rappresentazione latente con distribuzione marginale $p(z)$, soggetto al vincolo scalare $\mathbb{E}[\|z\|^2] = k$ (varianza unitaria per dimensione). Sia $\mathcal{R}_{lin}(p) = \inf_w \mathbb{E}[(w^T z - y)^2]$ il rischio di probing lineare ottimale per qualsiasi variabile target y statisticamente indipendente dalla chiave di probing w. Allora:

$\mathcal{R}_{lin}(p)$ è minimizzato — uniformemente su tutte le variabili target y — se e solo se $p = \mathcal{N}(0, I_k)$.

Dimostrazione per lo spazio di probing lineare. Sia il predittore lineare ottimale $w^* = \text{argmin}_w \mathbb{E}[(w^T z - y)^2] = \text{Cov}(y, z)\, [\text{Var}(z)]^{-1}$ (per la formula della regressione lineare). Il rischio ottimale è:

$\mathcal{R}_{lin}(p) = \text{Var}(y) - \text{Cov}(y, z)\, [\text{Cov}(z)]^{-1}\, \text{Cov}(z, y)$

Per massimizzare la riduzione del rischio $\text{Var}(y) - \mathcal{R}_{lin}(p) = \text{Cov}(y, z)\, [\text{Cov}(z)]^{-1}\, \text{Cov}(z, y)$, a parità di covarianza $\text{Cov}(y, z)$ tra target e rappresentazione, è necessario minimizzare $[\text{Cov}(z)]^{-1}$, ovvero massimizzare $\text{Cov}(z)$ nel senso dell'ordine di Loewner. Sotto il vincolo $\mathbb{E}[\|z\|^2] = k$ (equivalente a $\text{tr}(\text{Cov}(z)) = k$), la matrice di covarianza che massimizza la riduzione del rischio

uniformemente per qualsiasi direzione di probing è Cov(z) = I_k, la matrice identità — ovvero la distribuzione con varianze unitarie e correlazioni nulle tra dimensioni. La distribuzione a massima entropia soggetta a Cov(z) = I_k è la gaussiana standard isotropica $\mathcal{N}(0, I_k)$.

Per il probing non lineare, la dimostrazione si generalizza attraverso il principio del maximum mean discrepancy nello spazio di Hilbert a nucleo riproducente (RKHS) associato al kernel gaussiano $k(z, z') = \exp(-\|z - z'\|^2/2\sigma^2)$: le rappresentazioni con distribuzione $\mathcal{N}(0, I_k)$ minimizzano la MMD rispetto alla distribuzione ottimale per qualsiasi classificatore con kernel gaussiano, che per il teorema di universalità del kernel gaussiano approssima uniformemente qualsiasi funzione continua. □

B.2 — L'argomento modello-teorico di Putnam: costruzione esplicita

Enunciato (corollario del teorema di Löwenheim-Skolem e della completezza). Sia T una teoria del primo ordine coerente con almeno un modello infinito. Sia M l'interpretazione intenzionale di T (il «modello corretto»). Allora per ogni T e ogni M esistono infinite interpretazioni $M^* \neq M$ che soddisfano le stesse frasi di T e che assegnano referenti distinti ai termini di T.

Costruzione. Sia $\mathcal{D} = \mathrm{dom}(M)$ il dominio di M e sia P qualsiasi permutazione non triviale $\pi : \mathcal{D} \to \mathcal{D}$ (che esiste perché $\mathcal{D}$ è infinito). Si definisce M^* come segue: per ogni costante individuale c di T, $M^*(c) = \pi(M(c))$; per ogni predicato n-ario R di T, $M^*(R) = \{ (\pi(a_1), \ldots, \pi(a_n)) : (a_1, \ldots, a_n) \in M(R) \}$. Si dimostra per induzione sulla struttura delle formule che per ogni frase chiusa φ di T: $M \models \varphi$ se e solo se $M^* \models \varphi$. Caso base: $M \models R(c_1, \ldots, c_n)$ iff $(M(c_1), \ldots, M(c_n)) \in M(R)$ iff $(\pi(M(c_1)), \ldots, \pi(M(c_n))) \in M^*(R)$ iff $M^* \models R(c_1, \ldots, c_n)$. Il passo induttivo per connettivi e quantificatori è standard. Dunque M e M^* soddisfano esattamente le stesse frasi di T, ma $M^*(c) = \pi(M(c)) \neq M(c)$ per ogni costante c con $\pi(M(c)) \neq M(c)$. La struttura formale di T — le sue conseguenze logiche, le sue relazioni di dimostrazione — non distingue tra M e M^*. □

Corollario per i sistemi di apprendimento. Una rete neurale implementa una funzione $f_\theta : \mathcal{X} \to \mathcal{Y}$. Le sue «rappresentazioni» sono pattern di attivazione in strati intermedi. L'argomento di Putnam si applica: qualsiasi interpretazione delle rappresentazioni che soddisfa le relazioni formali apprese (le mappature input-output) è indistinguibile da qualsiasi altra interpretazione che soddisfi le stesse relazioni. Il sistema non può determinare a quale proprietà del mondo le sue rappresentazioni «si riferiscano» perché la sua struttura funzionale è compatibile con tutte le interpretazioni che preservano le mappature sul corpus di addestramento. □

B.3 — Il teorema di indefinibilità di Tarski: dimostrazione completa

Enunciato. Sia L un linguaggio formale del primo ordine sufficientemente espressivo da contenere l'aritmetica di Peano. Non esiste in L alcuna formula T(x) con una variabile libera tale che per ogni frase chiusa φ di L: $\vdash_L T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$, dove $\ulcorner \varphi \urcorner$ è il codice di Gödel di φ.

Dimostrazione. Si supponga per assurdo che tale T(x) esista. Per il lemma di diagonalizzazione (una conseguenza della capacità di L di rappresentare la funzione di sostituzione): esiste una frase λ di L tale che $\vdash_L \lambda \leftrightarrow \neg T(\ulcorner \lambda \urcorner)$. Si valuta la verità di λ nel modello standard $\mathbb{N}$.

Caso 1: λ è vera in $\mathbb{N}$. Per lo schema di Tarski (assunto come una condizione necessaria di adeguatezza di T): $T(\ulcorner \lambda \urcorner)$ è vera in $\mathbb{N}$. Ma $\lambda \leftrightarrow \neg T(\ulcorner \lambda \urcorner)$ implica che se λ è vera, allora $\neg T(\ulcorner \lambda \urcorner)$ è vera, ovvero $T(\ulcorner \lambda \urcorner)$ è falsa. Contraddizione.

Caso 2: λ è falsa in $\mathbb{N}$. Dalla falsità di λ e da $\lambda \leftrightarrow \neg T(\ulcorner \lambda \urcorner)$: $\neg T(\ulcorner \lambda \urcorner)$ è falsa, ovvero $T(\ulcorner \lambda \urcorner)$ è vera. Per lo schema di Tarski: λ è vera. Contraddizione.

Entrambi i casi producono contraddizione. Dunque T(x) non può esistere. □

Interpretazione. La verità semantica — la proprietà di essere vero nel modello standard — non può essere catturata da nessun predicato sintattico all'interno di L. La semantica eccede irreducibilmente la sintassi per qualsiasi linguaggio sufficientemente espressivo. □

B.4 — Il bound PAC-Bayes per encoder JEPA

Enunciato (forma semplificata). Sia $E_\theta : \mathscr{X} \to \mathbb{R}^k$ un encoder con parametri θ e sia $w \in \mathbb{R}^k$ un classificatore lineare. Siano P la distribuzione a priori sui parametri e Q la distribuzione a posteriori (empirica). Per ogni $\delta > 0$, con probabilità almeno $1 - \delta$ sull'estrazione del training set S di cardinalità m:

$$\mathcal{L}_{test}(\theta, w) \leq \mathcal{L}_{train}(\theta, w) + \sqrt{\left[\left(KL(Q \,\|\, P) + \ln(2\sqrt{m} / \delta)\right) / (2m) \right]}$$

Dimostrazione (schema). Il bound segue dall'applicazione del lemma di Donsker-Varadhan per la divergenza KL: per ogni funzione misurabile f e ogni distribuzione P, si ha $\mathbb{E}_Q[f] \leq KL(Q \,\|\, P) + \log \mathbb{E}_P[\exp(f)]$. Applicando questo lemma alla differenza $\mathcal{L}_{test} - \mathcal{L}_{train}$, mediando sulla distribuzione dei training set e applicando una diseguaglianza di concentrazione di Hoeffding per controllare le fluttuazioni della perdita empirica, si ottiene il bound enunciato. □

B.5 — Dinamica dell'EMA e prevenzione del collasso

Proposizione. Sia θ(t) il parametro del context encoder all'istante t e sia θ̄(t) il parametro del target encoder con aggiornamento EMA $\tau \in (0, 1)$. La differenza $\theta(t) - \bar{\theta}(t)$ soddisfa:

$\bar{\theta}(t) = (1 - \tau) \cdot \Sigma_{s=0}^{t} \tau^{t-s} \cdot \theta(s) + \tau^{t+1} \cdot \bar{\theta}(0)$

Dimostrazione. Svolgendo ricorsivamente l'aggiornamento $\bar{\theta}(t) = \tau \cdot \bar{\theta}(t-1) + (1 - \tau) \cdot \theta(t)$:

$\bar{\theta}(t) = \tau \cdot [\tau \cdot \bar{\theta}(t-2) + (1 - \tau) \cdot \theta(t-1)] + (1 - \tau) \cdot \theta(t) = \tau^2 \cdot \bar{\theta}(t-2) + (1 - \tau) \cdot \tau \cdot \theta(t-1) + (1 - \tau) \cdot \theta(t) = \ldots = \tau^{t+1} \cdot \bar{\theta}(0) + (1 - \tau) \cdot \Sigma_{s=0}^{t} \tau^{t-s} \cdot \theta(s)$. □

Corollario. $\bar{\theta}(t)$ è una media pesata esponenzialmente della storia di θ. La velocità di convergenza di $\bar{\theta}$ verso θ è dell'ordine di $(1 - \tau)$ / learning_rate. Per $\tau \to 1$, questa velocità tende a zero, rendendo il target encoder sempre più stabile rispetto alle variazioni del context encoder e prevenendo il collasso sincrono verso la soluzione degenere. □

Appendice C — Compendio delle Architetture

C.1 — Vision Transformer (ViT): varianti e parametri

La tabella seguente riassume le configurazioni standard del Vision Transformer usate nelle architetture JEPA.

ViT-S (Small): d = 384, teste MHSA = 6, layer L = 12, FFN dim = 1536, patch P = 16, parametri ≈ 22M.

ViT-B (Base): d = 768, teste MHSA = 12, layer L = 12, FFN dim = 3072, patch P = 16, parametri ≈ 86M.

ViT-L (Large): d = 1024, teste MHSA = 16, layer L = 24, FFN dim = 4096, patch P = 16, parametri ≈ 307M.

ViT-H (Huge): d = 1280, teste MHSA = 16, layer L = 32, FFN dim = 5120, patch P = 14, parametri ≈ 632M.

Per V-JEPA e V-JEPA 2, l'encoder processa cuboid 3D $P \times P \times \tau$ con $\tau = 2$ frame; il numero di token per clip di T = 16 frame con risoluzione 224×224 è $N_{st} = (16/2) \times (224/16)^2 = 8 \times 196 = 1568$.

C.2 — Recurrent State Space Model (RSSM): formalismo completo

Il RSSM (Hafner et al., 2019, 2020, 2021) fattorizza la rappresentazione latente in una componente deterministica h_t e una stocastica z_t:

Componente deterministica: $h_t = f_\theta(h_{t-1}, z_{t-1}, a_{t-1})$ dove f_θ è tipicamente una GRU.

Componente stocastica (prior): $z_t \sim p_\theta(z_t \mid h_t) = \mathcal{N}(\mu_\theta(h_t), \sigma^2_\theta(h_t))$ per DreamerV1/V2 (latenti gaussiani), oppure $z_t \sim \mathrm{Cat}(p_\theta(h_t))$ per DreamerV2/V3 (latenti categorici con 32 categorie di 32 classi ciascuna).

Componente stocastica (posteriore): $z_t \sim q_\varphi(z_t \mid h_t, o_t) = \mathcal{N}(\mu_\varphi(h_t, o_t), \sigma^2_\varphi(h_t, o_t))$ per latenti gaussiani, oppure $q_\varphi(z_t \mid h_t, o_t) = \mathrm{Cat}(p_\varphi(h_t, o_t))$ per latenti categorici.

Il modello di osservazione: $o_t \sim p_\theta(o_t \mid h_t, z_t)$.

La perdita ELBO (Evidence Lower Bound):

$$\mathcal{L}_{\mathrm{RSSM}} = \mathbb{E}_q \left[-\log p_\theta(o_t \mid h_t, z_t) + \beta \cdot \mathrm{KL}(q_\varphi(z_t \mid h_t, o_t) \,\|\, p_\theta(z_t \mid h_t)) \right]$$

In DreamerV2, il KL balancing sostituisce la minimizzazione del KL completo con una media pesata tra due obiettivi parziali:

$$\mathcal{L}_{\mathrm{KL}} = (1 - \alpha) \cdot \mathrm{KL}(\mathrm{sg}(q_\varphi) \,\|\, p_\theta) + \alpha \cdot \mathrm{KL}(q_\varphi \,\|\, \mathrm{sg}(p_\theta))$$

dove $\alpha \in [0, 1]$ e il parametro sg blocca il gradiente. Con $\alpha = 0.8$ si ottiene empiricamente la migliore stabilità di addestramento.

C.3 — HiFi-GAN: struttura del decoder

HiFi-GAN (Kong et al., 2020) è un generatore adversariale per la sintesi di forma d'onda audio da feature di mel-spectrogram o da feature latenti compresse. Il generatore ha struttura multi-scala con moduli di upsampling e residual blocks:

Struttura base: feature input → conv 7×1 → [upsampling + Multi-Receptive-Field Fusion (MRF)] × n_ups → conv → Tanh → forma d'onda.

Ciascun modulo MRF è una somma di k blocchi residuali con receptive fields di dimensioni diverse (tipicamente k = 3 con dilated conv di dilation = (1, 2, 4), (2, 4, 8), (4, 8, 16)) che permette al decoder di catturare strutture di periodicità a scale temporali multiple.

Nel sistema JEPA+DAAM del Capitolo VII, il decoder HiFi-GAN modifica le operazioni di normalizzazione per ricevere i gate DAAM come condizionamento: le Adaptive Instance Normalization (AdaIN) vengono sostituite con Gated Instance Normalization (GIN) che scala e trasla le attivazioni in funzione dei gate $G(\mathbf{x}_t)$ appresi dall'encoder.

Appendice D — Glossario Terminologico

Apprendimento autosupervisionato (self-supervised learning, SSL). Paradigma di addestramento in cui il segnale supervisorio è derivato automaticamente dai dati non etichettati attraverso compiti ausiliari

(completamento, contrastività, predizione) senza richiedere annotazioni umane esplicite.

Catastrophic forgetting (oblio catastrofico). Fenomeno per cui l'apprendimento su nuovi dati sovrascrive i parametri che codificano l'apprendimento su dati precedenti, portando a un marcato degrado delle prestazioni sui compiti appresi in precedenza.

Cecità di secondo ordine. In epistemologia dei paradigmi scientifici, il fallimento di riconoscere che si sta fallendo di riconoscere qualcosa — l'invisibilità della propria invisibilità critica. Si manifesta quando i meccanismi di conversione delle anomalie in virtù operano al livello del vocabolario di descrizione prima che al livello del ragionamento esplicito.

Collasso delle rappresentazioni (representation collapse). Degenerazione dell'addestramento di encoder in reti JEPA o contrastive in cui tutte le rappresentazioni convergono verso un valore costante (collasso totale) o verso uno spazio di dimensione molto inferiore a quella nominale (collasso parziale), eliminando la discriminabilità delle feature.

Curazione dei dati (data curation). Processo di selezione e filtro del corpus di addestramento in funzione di un criterio di qualità o di distribuzione bersaglio. In V-JEPA 2, la curazione avviene per prossimità di embedding CLIP rispetto ai dataset di benchmark, introducendo circolarità tra la distribuzione di addestramento e quella di valutazione.

Disentanglement. Proprietà di una rappresentazione latente in cui le dimensioni individuali corrispondono a fattori causali indipendenti del mondo (forma, colore, posizione, identità), permettendo la manipolazione indipendente di ciascun fattore cambiando soltanto la dimensione corrispondente nella rappresentazione.

Distribution shift (spostamento di distribuzione). Divergenza tra la distribuzione dei dati di addestramento e quella incontrata all'inferenza. In V-JEPA 2, lo spostamento di distribuzione produce degradazione marcata delle prestazioni su compiti fisicamente equivalenti ma con caratteristiche sensoriali non viste durante l'addestramento.

EMA (Exponential Moving Average, media mobile esponenziale). Aggiornamento del target encoder $\bar{\theta} \leftarrow \tau \cdot \bar{\theta} + (1 - \tau) \cdot \theta$ con $\tau \in (0, 1)$, che produce parametri ritardati rispetto al context encoder e previene il collasso sincrono verso soluzioni degeneri.

Embedding posizionale. Rappresentazione della posizione spaziale (o spaziotemporale) di una patch all'interno dell'immagine (o del video), tipicamente come vettore sinusoidale o come parametro apprendibile, aggiunta alla proiezione lineare della patch prima del processing Transformer.

Frame problem. In intelligenza artificiale, il problema della pertinenza: come un sistema possa determinare quali azioni e quali aspetti del mondo sono rilevanti per

un dato obiettivo senza esaminare esaustivamente tutte le possibilità. Nei sistemi JEPA, si manifesta come dipendenza dalla curazione esterna dei dati: il sistema non può determinare dall'interno delle proprie operazioni quali strutture percettive siano semanticamente rilevanti.

GELU (Gaussian Error Linear Unit). Funzione di attivazione GELU(x) = x · Φ(x) dove Φ è la CDF della distribuzione normale standard. Approssimazione comune: GELU(x) ≈ 0.5 · x · (1 + tanh(√(2/π) · (x + 0.044715 · x³))).

Geodetica latente. Curva di lunghezza minima che connette due punti in una varietà latente dotata di metrica riemanniana. Nella Semantic Tube Prediction (STP), le traiettorie dei token a convergenza seguono geodetiche nel manifold latente, il che equivale a equazioni differenziali del tipo $\ddot{h}_t + \Gamma^k_{ij}(h_t)\, \dot{h}^i_t\, \dot{h}^j_t = 0$ dove Γ sono i simboli di Christoffel della metrica appresa.

Information Bottleneck. Principio di apprendimento delle rappresentazioni (Tishby et al., 1999) che minimizza l'informazione sull'input I(X; Z) preservata dalla rappresentazione Z, massimizzando simultaneamente l'informazione sul target I(Z; Y): $\min_{p(z|x)} [-I(Z; Y) + \beta \cdot I(X; Z)]$.

DaM-GMJE (Data as Mode GMJE). Formulazione non-parametrica di GMJE in cui ogni coppia di embedding storici nella memory bank viene trattata come un componente equiprobabile della mistura congiunta. Conduce algebricamente alla perdita InfoNCE quando la covarianza viene ristretta a un valore isotropico block-diagonale con blocchi off-diagonali nulli ($\Sigma_{ct} = 0$).

GJE (Gaussian Joint Embeddings). Framework probabilistico per JEPA introdotto da Huang (2026) che sostituisce la predizione deterministica $z_t = g(z_c)$ con la modellazione della densità congiunta $p(z_c, z_t)$. Esiste in forma duale (Gaussian Process nel sample space, complessità $O(N^3)$) e primale (feature covariance nel feature space, complessità $O(d^3)$). La formulazione primale offre la predizione ottimale $z^*_t = C_{tc} C_{cc}^{-1} z^*_c$ in forma chiusa.

GMJE (Gaussian Mixture Joint Embeddings). Estensione multimodale di GJE basata su misture di K gaussiane congiunti. Risolve la limitazione unimodale di GJE permettendo al predittore di instradare predizioni verso modi semantici distinti attraverso pesi di routing data-dipendenti $\gamma_k(z_c)$. Le varianti architetturali includono GMJE-Proto (prototipi globali fissi), GMJE-MDN (parametri dinamici condizionati su z_c), GMJE-GNG (scoperta topologica adattiva) e SMC-GMJE (particle filter per la memory bank).

GMJE-GNG (GMJE con Growing Neural Gas). Variante di GMJE che scopre automaticamente il numero K di componenti semantici e la topologia del manifold congiunto attraverso l'algoritmo di Growing Neural Gas. Inizia con due prototipi e inserisce nuovi nodi nelle regioni con errore di quantizzazione elevato. Rimuove connessioni obsolete tramite invecchiamento degli archi. Non richiede la specificazione di K a priori.

GMJE-MDN (GMJE con Mixture Density Network). Variante di GMJE in cui i parametri della mistura (μ_k, Σ_k, π_k) sono funzioni dinamiche del contesto z_c predette da una rete neurale parametrica. Il bottleneck informativo (la rete osserva solo z_c, non [z_c; z_t]) previene l'identity collapse. Produce la distribuzione condizionale più espressiva della famiglia GMJE, adattando la forma della distribuzione all'ambiguità specifica di ciascun input.

InfoNCE bridge. Risultato teorico di Huang (2026) che dimostra formalmente che la perdita InfoNCE — usata in SimCLR, MoCo e tutti i principali metodi contrastivi — è un caso limite degenere di GMJE, ottenuto imponendo covarianza isotropica block-diagonale (Σ_ct = 0), pesi uniformi (π_m = 1/M) ed embedding normalizzati. Unifica le famiglie contrastiva e predittiva dell'SSL sotto il medesimo principio generativo.

Mahalanobis Trace Trap. Patologia algebrica identificata da Huang (2026) nell'ottimizzazione diretta della verosimiglianza congiunta GJE: il termine di data-fit (1/N) Σ_i z_i^T C^{-1} z_i si annulla identicamente a d quando C è la covarianza empirica calcolata dallo stesso batch Z. Il gradiente è zero: nessun segnale attrattivo cross-view sopravvive. L'ottimizzatore collassa le rappresentazioni verso l'origine minimizzando il solo termine log|C|. Le varianti architetturali GMJE-Proto, MDN e SMC-GMJE disaccoppiano i parametri di covarianza dal batch corrente per aggirare questo annullamento.

SMC-GMJE (Sequential Monte Carlo GMJE). Variante non-parametrica di GMJE che sostituisce la memory bank FIFO dei metodi contrastivi con un particle filter probabilistico. Ogni particella (embedding storico) ha un peso di importanza W^(m) aggiornato ricorsivamente in base alla sua verosimiglianza sotto il batch corrente. Il Sequential Importance Resampling clona le particelle con alto peso (hard negatives informativi) e elimina quelle con basso peso (campioni ridondanti). Su CIFAR-10 con M=256, ottiene +3.91% di linear probing accuracy rispetto a MoCo v2 FIFO.

JEPA (Joint-Embedding Predictive Architecture). Famiglia di architetture autosupervisionato in cui il compito di apprendimento consiste nel predire la rappresentazione latente di una parte dell'input (target) a partire dalla rappresentazione latente di un'altra parte (contesto), senza ricostruzione nello spazio delle osservazioni.

Lavaggio epistemico (epistemic laundering). Meccanismo sistematico per cui struttura ereditata entra in un sistema in un punto, scompare dalla visibilità durante l'elaborazione opaca, e riemerge all'output attribuita erroneamente al sistema come scoperta emergente o capacità genuinamente acquisita.

Layer Normalization (LN). Normalizzazione delle attivazioni $LN(x) = (x - \mu) / \sigma \cdot \gamma + \beta$ dove μ e σ sono media e deviazione standard calcolate sulle dimensioni del layer per ciascun campione, e γ, β sono parametri apprendibili.

Nucleo inmetabolizzabile. Assunzione costitutiva di un paradigma scientifico che non può essere messa in questione dall'interno del paradigma senza dissolverlo. Per il paradigma computazionale dell'IA, il nucleo inmetabolizzabile è la tesi che la cognizione sia esauribile attraverso il calcolo formale.

Normatività. Proprietà delle regole e degli standard di specificare come le cose dovrebbero essere, non soltanto come sono. Comprendere una regola richiede normatività: sapere che un'applicazione è corretta o scorretta indipendentemente dalla frequenza con cui viene eseguita.

Paradigma meccanicistico. Sistema di presupposizioni condiviso dalla fisica teorica moderna e dall'intelligenza artificiale contemporanea: fondamenta discrete, operazioni formali, aspirazione all'indipendenza dall'osservatore, composizione come principio esplicativo, meccanismi come forma della spiegazione.

Seam paradigmatica (cucitura paradigmatica). Punto di una struttura concettuale in cui le assunzioni del paradigma generano conseguenze che contraddicono quelle stesse assunzioni. Si distingue da un problema non risolto perché non è risolvibile attraverso risorse interne al paradigma: richiede la revisione di assunzioni fondamentali.

SIGReg (Stochastic Isotropic Gaussian Regularization). Regolarizzatore di LeJEPA che impone, attraverso proiezioni stocastiche su iperpiani casuali e test di Epps-Pulley, che la distribuzione marginale degli embedding sia gaussiana isotropica, garantendo la distribuzione ottimale per qualsiasi compito di probing dimostrata dal Teorema 1.

Softmax. Funzione $\mathrm{softmax}(x)_i = \exp(x_i) / \Sigma_j \exp(x_j)$ che trasforma un vettore di logit in una distribuzione di probabilità. Nell'attenzione Transformer: $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(QK^T / \sqrt{d_k}) \cdot V$.

STP (Semantic Tube Prediction). Obiettivo di regolarizzazione per LLM che aggiunge alla cross-entropy standard un termine di penalizzazione della distanza euclidea tra lo stato nascosto h_t e la sua proiezione sulla retta (geodetica) che connette h_{t-K} e h_{t+K} nel spazio latente, migliorando il rapporto segnale-rumore delle traiettorie latenti.

Stop-gradient (sg). Operazione che propaga il valore in avanti ma blocca il flusso del gradiente in backpropagation: $\mathrm{sg}(x) = x$, $\partial \mathrm{sg}(x)/\partial x = 0$. Usata per impedire al gradiente di fluire attraverso il target encoder, preservando la natura del target come «supervisore» stabile.

Symbol Grounding Problem. Problema filosofico formulato da Harnad (1990): come possono i simboli manipolati da un sistema formale acquistare significato (essere «radicati» nel mondo) per il sistema stesso, piuttosto che soltanto per gli osservatori esterni che interpretano il sistema?

ViT (Vision Transformer). Architettura transformer applicata a immagini mediante la suddivisione in patch non sovrapposte, la proiezione lineare di

ciascuna patch in un embedding vettoriale e l'applicazione di blocchi di self-attention multi-testa.

Bibliografia

A — Architetture e apprendimento autosupervisionato

Huang, Y. (2026). Gaussian Joint Embeddings For Self-Supervised Representation Learning. arXiv preprint arXiv:2603.26799v1. [Introduce GJE/GMJE — il framework probabilistico per JEPA basato su densità congiunta. Dimostra algebricamente il Mahalanobis Trace Trap, unifica contrastive learning e SSL predittivo nell'InfoNCE bridge, introduce SMC-GMJE come memory bank ottimale e stabilisce la capacità generativa del latent manifold.]

Kuang, Y., Dagade, Y., Rudner, T. G. J., Balestriero, R., & LeCun, Y. (2026). Rectified LpJEPA: Joint-Embedding Predictive Architectures with Sparse and Maximum-Entropy Representations. arXiv preprint arXiv:2602.01456. [Introduce RDMReg e la famiglia RGG. Dimostra che la distribuzione a massima entropia sotto vincoli di norma ℓp con rettificazione è la Gaussiana Generalizzata Rettificata, con sparsità ℓ0 controllabile analiticamente tramite μ. Rectified LpJEPA generalizza strettamente LeJEPA e ottiene tradeoff sparsità–prestazione favorevoli su ImageNet-100.]

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., & Ballas, N. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15619–15629.

Assran, M., et al. (2025). V-JEPA 2: Self-Supervised Video Understanding. *Meta AI Research Technical Report*.

Bardes, A., Ponce, J., & LeCun, Y. (2022). VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. *International Conference on Learning Representations (ICLR)*.

Bao, H., Dong, L., Piao, S., & Wei, F. (2022). BEiT: BERT Pre-Training of Image Transformers. *International Conference on Learning Representations (ICLR)*.

Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *International Conference on Machine Learning (ICML)*, 1597–1607.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660.

Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2933–2941.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Dong, Y., Cordonnier, J.-B., & Loukas, A. (2021). Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth. *International Conference on Machine Learning (ICML)*, 2793–2803.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*.

Grill, J.-B., Strub, F., Altché, F., et al. (2020). Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 21271–21284.

Ha, D., & Schmidhuber, J. (2018). World Models. *arXiv preprint arXiv:1803.10122*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Huang, A., LeCun, Y., & Balestriero, R. (2026). Semantic Tube Prediction: Geometric Regularization for Efficient Fine-Tuning of Large Language Models. *arXiv preprint (2026)*.

Ioannides, A., et al. (2025). JEPA as Neural Tokenizer for Speech: Adaptive Density Attention and Finite Scalar Quantization in a Two-Stage Self-Supervised Framework. *arXiv preprint (2025)*.

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 17022–17033.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. *OpenReview preprint*. Disponibile a: https://openreview.net/pdf?id=BZ5a1r-kVsf

LeCun, Y. (2024). Intervista, Lex Fridman Podcast #416. Marzo 2024. Trascrizione: https://lexfridman.com/yann-lecun-3-transcript/

Nguyen, T., et al. (2024). LeJEPA: Learning Joint Embeddings with Provably Optimal Representation Distribution. *arXiv preprint (2024/2025)*.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.

Oquab, M., Darcet, T., Moutakanni, T., et al. (2024). DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research (TMLR)*.

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning (ICML)*, 8748–8763.

Shorten, C., & Khoshgoftaar, T. M. (2019). A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), 60.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The Information Bottleneck Method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computation*, 368–377.

B — World Models e apprendimento per rinforzo basato su modello

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). Dream to Control: Learning Behaviors by Latent Imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D., Lillicrap, T., Norouzi, M., & Ba, J. (2020). Mastering Atari with Discrete World Models. *International Conference on Learning Representations (ICLR)*.

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2023). Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104*.

Hu, A., Russell, L., Hudson, B., et al. (2023). GAIA-1: A Generative World Model for Autonomous Driving. *arXiv preprint arXiv:2309.17080*.

Lecun, Y. (2022). Op. cit. [cfr. sezione A].

Matsuo, Y., et al. (2022). Deep Learning, Reinforcement Learning, and World Models. *Neural Networks*, 152, 267–291.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2ª ed.). MIT Press.

C — Scalabilità, Scaling Laws e sostenibilità computazionale

Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., & Sutskever, I. (2018). AI and Compute. *Blog post, OpenAI*. Disponibile a: https://openai.com/research/ai-and-compute

Brown, T., et al. (2020). Op. cit. [cfr. sezione A].

Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 30016–30030. (Paper «Chinchilla».)

Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research (TMLR)*.

D — Filosofia della mente e dell'intelligenza artificiale

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Brooks, R. A. (1991). Intelligence Without Representation. *Artificial Intelligence*, 47(1–3), 139–159.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.

Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row.

Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard. [Trad. it.: *Fenomenologia della percezione*, Bompiani, 2003.]

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. *Minnesota Studies in the Philosophy of Science*, 1, 253–329.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

E — Filosofia del linguaggio e critica della tesi computazionale

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50. [Trad. it.: «Senso e denotazione», in *Logica e aritmetica*, Boringhieri, 1965.]

Hamann, J. G. (1784). Metakritik über den Purismus der Vernunft. In K. Haynes (ed. e trad.), *Hamann: Writings on Philosophy and Language*, Cambridge University Press, 2007.

Kripke, S. A. (1982). *Wittgenstein on Rules and Private Language*. Harvard University Press.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge University Press.

Vygotsky, L. S. (1934/1986). *Thought and Language* (A. Kozulin, trad.). MIT Press.

Wittgenstein, L. (1953). *Philosophische Untersuchungen / Philosophical Investigations* (G. E. M. Anscombe, trad.). Blackwell. [Trad. it.: *Ricerche filosofiche*, Einaudi, 2014.]

F — Epistemologia, filosofia della scienza e teoria dei paradigmi

Bedau, M. A. (1997). Weak Emergence. *Philosophical Perspectives*, 11, 375–399.

Gödel, K. (1930). Die Vollständigkeit der Axiome des logischen Funktionenkalküls. *Monatshefte für Mathematik und Physik*, 37, 349–360.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38, 173–198. [Trad. it. in *Opere*, Bollati Boringhieri, 1999.]

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press. [Trad. it.: *La struttura delle rivoluzioni scientifiche*, Einaudi, 1969.]

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes: Philosophical Papers, Volume 1* (J. Worrall & G. Currie, eds.). Cambridge University Press.

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

Tarski, A. (1936). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica*, 1, 261–405. [Trad. it. in *Logica, semantica, metamatematica*, Il Saggiatore, 1969.]

G — Fisica teorica e teoria delle stringhe

Green, M. B., & Schwarz, J. H. (1984). Anomaly Cancellations in Supersymmetric D=10 Gauge Theory and Superstring Theory. *Physics Letters B*, 149(1–3), 117–122.

Greene, B., & Witten, E. (2025). String Theory 2037: A Conversation. *World Science Festival*, December 2025.

Hořava, P., & Witten, E. (1996). Heterotic and Type I String Dynamics from Eleven Dimensions. *Nuclear Physics B*, 460(3), 506–524.

Maldacena, J. (1998). The Large N Limit of Superconformal Field Theories and Supergravity. *Advances in Theoretical and Mathematical Physics*, 2(2), 231–252.

Penrose, R. (2004). *The Road to Reality: A Complete Guide to the Laws of the Universe*. Jonathan Cape.

Susskind, L. (2005). *The Cosmic Landscape: String Theory and the Illusion of Intelligent Design*. Little, Brown.

Weinberg, S. (1987). Anthropic Bound on the Cosmological Constant. *Physical Review Letters*, 59(22), 2607–2610.

Witten, E. (1988). Topological Quantum Field Theory. *Communications in Mathematical Physics*, 117(3), 353–386.

Witten, E. (1995). String Theory Dynamics in Various Dimensions. *Nuclear Physics B*, 443(1–2), 85–126.

H — Letteratura critica integrativa

Nourizadeh, M. (2026). Syntax Without Semantics: V-JEPA 2 and String Theory as Isomorphic Failures of the Mechanistic Paradigm. *Zenodo*, DOI: 10.5281/zenodo.18236574. (Version 1.2, January 2026.)

I — Riferimenti per lo sviluppo del parlato e della psicologia cognitiva

Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive Field Effects. *Nature Neuroscience*, 2(1), 79–87.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.